Advances on Semantic Web and New Technologies

March, 2009

Editors:

Dra. María Josefa Somodevilla García

Dr. David Eduardo Pinto Avendaño

The Workshop on Semantic Web and New Technologies was held by second time at the Faculty of Computer Science of Benemérita Universidad Autónoma de Puebla, Mexico in March 2009.

The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries. Semantic Web technologies are beginning to play a significant role in many diverse areas, marking a turning point in the evolution of the Web.

The goal of this workshop is to provide a forum for the Semantic Web community, in which participants can present and discuss approaches to add semantics on the Web, show innovative applications in this field and identify upcoming research issues related to Semantic Web. In order to fulfill these objectives, the more important workshop topics included Semantic Search, Semantic Advertising and Marketing, Linked Data, Collaboration and Social Network, Foundational Topics, Semantic Web and Web 3.0, Ontologies, Semantic Integration, Data Integration and Mashups, Unstructured Information, Semantic Query, Semantic Rules, Developing Semantic Applications and Semantic SOA.

Dr John Cardiff was the invited speaker in this Second Workshop on Semantic Web, who is a fulltime lecturer and lead researcher in the Social Media Research Group, based at the Institute of Technology Tallaght, Dublin, Ireland. He has previously held positions in the Department of Computer Science, Trinity College Dublin, and in the University of Queensland, Australia, where he obtained the degree of Ph.D. in 1990. He has extensive experience in semantic web technologies, heterogeneous database research and query processing and optimization. He collaborates closely with researchers of the National Language Engineering Laboratory at the Polytechnic University of Valencia, Spain, the Knowledge and Data Engineering Group of Trinity College Dublin, and the IBM Dublin Center for Advanced Studies. He is currently supervising four PhD students who are investigating semantic web based recommender systems, blogosphere analysis, and adaptive hypermedia systems. Dr Cardiff has a wide breadth of experience of research and management of large European Union funded projects under programmes such as RACE, Esprit, and AIM. He has over 20 refereed publications in international conferences and journals.

Content

Content

The Evolution of the Semantic Web

John Cardiff

Social Media Research Group,
Institute of Technology Tallaght, Dublin, Ireland
email John.Cardiff@ittdublin.ie

Abstract — **The Semantic Web offers an exciting promise of a world in which computers and humans can cooperate effectively with a common understanding of the meaning of data. However, in the decade since the term has come into widespread usage, Semantic Web applications have been slow to emerge from the research laboratories. In this paper, we present a brief overview of the Semantic Web vision and the underlying technologies. We describe the advances made in recent years and explain why we believe that Semantic Web technology will be the driving force behind the next generation of Web applications.**

Keywords: Semantic Web, Ontology, Web of Data

I. INTRODUCTION

The World Wide Web (WWW) was invented by Tim Berners Lee in 1989, while he was working at the European Laboratory for Particle Physics (CERN) in Switzerland. It was conceived as a means to allow physicists working in different countries to communicate and to share documentation more efficiently. He wrote the first browser and Web server, allowing hypertext documents to be stored, retrieved and viewed.

The Web added two important services to the internet - it provided a very convenient means for us to retrieve and view information - we can then see the web as a vast document store in which we retrieve documents (web pages) by typing in their address into a web browser. Secondly, it provided a language called HTML, which describes to computers how to display documents written in this language. Documents, or web pages, are accessed by a unique identifier called a Uniform Resource Locator (URL) and are accessed using a Web browser. Within a short space of time, the WWW had become a popular infrastructure for sharing information, and as the volume of information increased its use became increasingly widespread.

Although the web provides the infrastructure for us to publish and retrieve documents, the HTML language defines only the visual characteristics, ie. how the documents are to be presented on a computer screen to the user. It is up to the user who requested the document to interpret the information it contains. This seems counterintuitive, as we normally think of computers as the tools to perform the more complex tasks, making life easier for humans. The problem is that within HTML there is no consideration of the meaning of the document, they are not represented in a way that allows interpretation of their information content by computers.

If computers could interpret the content of a web page, a lot of exciting possibilities would arise. Information could be exchanged between machines, automated processing and integration of data on different sites could occur. Fundamentally, they could improve the ways in which they can retrieve and utilise the information for us because they would have an understanding of what we are interested in. This is where the Semantic Web fits into the picture - today's web (the "syntactic" web) is about documents whereas the semantic web is about "things" - concepts we are interested in (people, places, events etc.), and the relationships between these concepts.

The Semantic Web vision envisages advanced management of the information on the internet, allowing us to pose queries rather than browse documents, to infer new knowledge from existing facts, and to identify inconsistencies. Some of the advantages of achieving this goal include [4]:

- The ability to locate information based on its meaning, eg. knowing when two statements are equivalent, or knowing that a reference to a person in different web pages are referring to the same individual.
- Integrating information across different sources − by creating mappings across application and terminological boundaries we can identify identical or related concepts,
- Improving the way in which information is presented to a user, eg. aggregating information from different sources, removing duplicates, and summarising the data.

While the technologies to enable the development of the Semantic Web were in place from the conception of the web, a seminal article by Tim Berners-Lee, James Hendler and Ora Lassila [1] published in *Scientific American* in 2001 provided the impetus for research and development to commence. The authors described a world in which independent applications could cooperate and share their data in a seamless way to allow the user to achieve a task with minimal intervention. Central to this vision is the ability to "unlock" data that is controlled by different applications and make it available for use by other applications. Much of this data is already available on the

Web, for example we can access our bank statements, our diaries and our photos online. But the data is controlled by proprietary applications. The Semantic Web vision is to publish this data in a sharable form − we could integrate the items of our bank statements into our calendar so that we could see what transactions we made on that day, or include photos so that we could see what we were doing at that time.

However, eight years after publication of this article, we are still some distance realising this vision. In this paper, present an overview of the Semantic Web. We explain why progress has been slow and the reasons we believe this to be about to change.

The paper is organized as follows. In Section II we describe the problems we face when trying to extract meaning from the web as it is today. Section III presents a brief overview of the technologies underlying the Semantic Web. In Section IV we give an overview of the gamut of typical Semantic Web applications and Section V introduces the Linking Open Data project. Finally, we present our conclusions and look to the future in Section VI.

II. THE PROBLEM WITH THE "SYNTACTIC WEB"

In Figure 1 we see a "typical" web page written in HTML which we will use to exemplify some of the drawbacks of the traditional web. This page lists the keynote speeches which took place at the 2009 World Wide Web conference[1]. To the reader, the content of the page can be interpreted intuitively. We can read the titles of the speeches, the names of the speakers and the time and dates at which they take place. Furthermore, by familiarity with browser interaction paradigms, we can realize that by following a hyperlink we can retrieve information about concepts related to the conference (authors, sponsors, attendees etc.). In this example, by following the hyperlink labelled "Sir Tim Berners-Lee" we will retrieve a document containing information about the person of this name. We intuitively assign a meaning - perhaps "has-homepage" - to the hyperlink, allowing us to assimilate the information presented to us.

A web browser cannot assign any to these links we see in this page − a hyperlink is simply a link from one document to another and the interpretation of the meaning of the link (and of the documents themselves!) is a task for the human reader. All that can be inferred automatically is that some undefined association between the two documents exists.

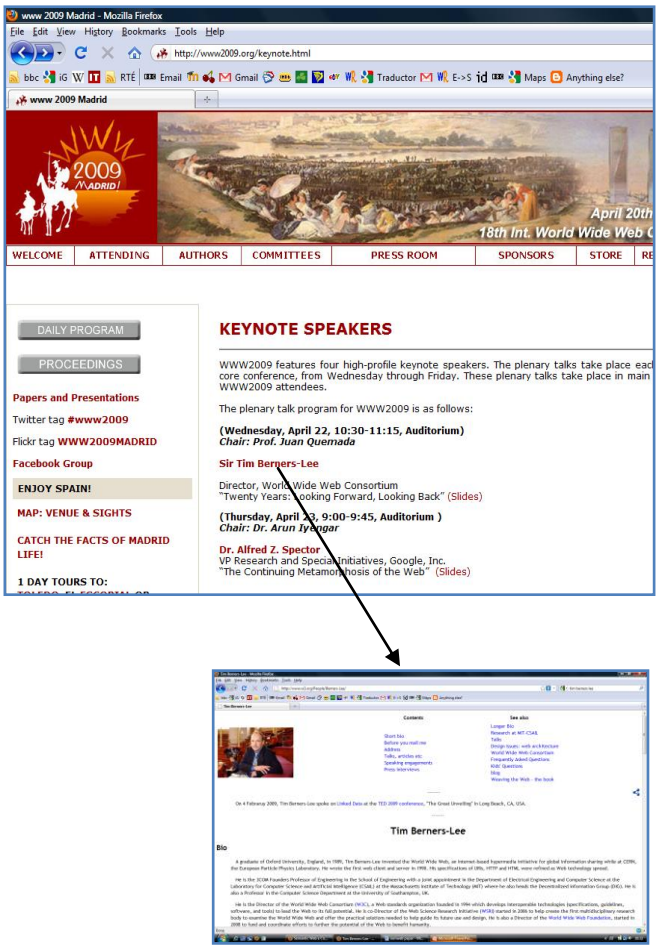


Figure 1. "Traditional" Web Pages with hyperlinks

The problems are even more clear when we consider the nature of keyword-based browsing. While search engines such as Google and Yahoo! are clearly very good at what they do, we frequently are presented with a vast number of results, many (most?) of which will be irrelevant to our search. Semantically similar items will not be retrieved (for instance a search for "movie" will not retrieve results where the word "film" was used). And most significantly, the result set is a collection of individual web pages. Our tasks often require access to multiple sites (such as when we book a holiday), and so it is our responsibility to formulate a sequence of queries to retrieve the individual web pages, each one of which performs part of the task at hand.

There are two potential ways to deal with this problem. One approach is to take the web as it is currently implemented, and to use Artificial Intelligence techniques to analyze the content of web pages in order to provide an interpretation of its meaning. This approach however would be prone to error and would require validation. Furthermore, the rate at which the web is growing would render it practically impossible to achieve.

The other approach is to represent the web pages in a form in which we can represent and interpret the data they contain. If there is a common representation to express the

[1] http://www2009.org/keynote.html

meaning of the data on the web, we can then develop languages, reasoners, and applications which can exploit this representation. This is the approach of the Semantic Web.

III. SEMANTIC WEB TECHNOLOGIES

The Semantic Web describes a web of data rather than documents. And just as we need common formats and standards to be able to retrieve documents from computers all over the world, we need common formats for the representation and integration of data. We also need languages that allow us to describe how this data relates to real world objects and to reason about the data. The famous "Layer Cake" [10] diagram, shown in Figure 2, gives an overview of the hierarchy of the principal languages and technologies, each one exploiting the features of the levels beneath it. It also reinforces the fact that the Semantic Web is not separate from the existing web, but is in fact an extension of its capabilities.

In this section, we summarize and discuss the key aspects shown in the Layer Cake diagram. Firstly we describe the core technologies: the languages RDF and RDFS. Next we describe the higher level concepts, focusing in particular on the concept of the ontology which is at the heart of the Semantic Web infrastructure. Finally we examine the trends and directions of the technology. For further information on the concepts presented in this section, the reader is referred to a more detailed work (eg. [4], [5]).

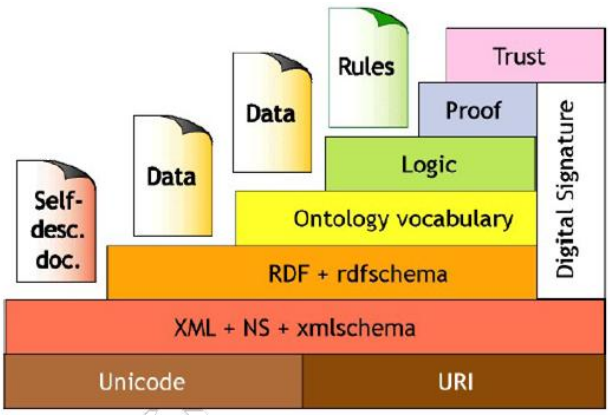


Figure 2. The Semantic Web Layers

A. The Core Technologies: RDF and RDFS

What HTML is to documents, RDF (Resource Description Framework) is to data. It is a W3C standard[2] based on XML which allows us to make statements about objects. It is a data model rather than a language - we can say that an object possesses a particular property, or that it has a named relationship with another object. RDF statements are written as triples: a subject, predicate and object.

By way of example, the statement

"The Adventures of Tom Sawyer" was written by Mark Twain

could be expressed in RDF by a statement such as

```
<rdf:Description
      rdf:about=www.famouswriters.org/twain/mark>
                <s:hasName>Mark Twain</s:hasName>
                <s:hasWritten rdf:resource=
                        www.books.org/ISBN0001047>
</rdf:Description>
```

At first glance it may appear that this information could be equally well represented using XML. However XML makes no commitment on which words should be used to describe a given set of concepts. In the above example we have a property entitled "*hasWritten*", but this could equally have been "*IsAuthorOf*" or another such variant. So, XML is suitable for closed and stable domains, rather than for sharable web resources.

The statements we make in RDF are unambiguous and have a uniform structure. Concepts are each identified by a Universal Resource Identifer (URI) which allows us to make statements about the same concept in different applications. This provides the basis for semantic interoperability, allowing us to distinguish between ambiguous terms (for instance an address could be a geographical location, or a speech) and to define a place on the web at which we can find the definition of the concept.

To describe and make general statements collectively about groups of objects (or classes), and to assign properties to members of these groups we use RDF Schema, or RDFS[3]. RDFS provides a basic object model, and enables us to describe resources in terms of classes, properties, and values. Whereas in RDF we spoke about specific objects such as "“The Adventures of Tom Sawyer” and "Mark Twain", in RDFS we can make general statements such as

"A book was written by an author"

This could be expressed in RDFS as

```
<rdf:Property rdf:ID=“HasWritten”
       <rdfs:domain rdf:resource=“#author”\>
       <rdfs:range rdf:resource=“#book”\>
<\rdf:Property>
```

An expansion of these examples, and the relationship between the graphical representations of RDF and RDFS is shown in Figure 3.

[2] www.w3.org/RDF/

[3] http://www.w3.org/TR/rdf-schema/

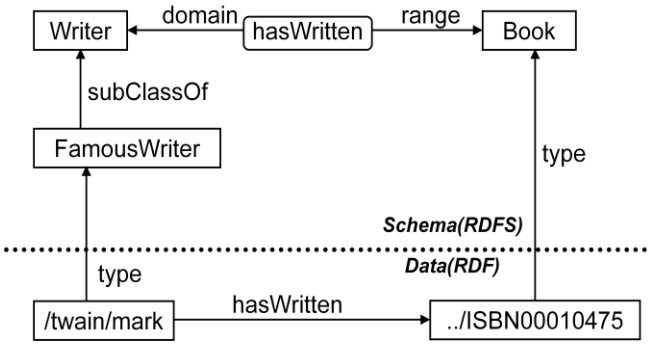


Figure 3. Relationship between RDF and RDFS [5]

B. Ontologies and Reasoning

RDF and RDFS allow us to describe aspects of a domain, but the modelling primitives are too restrictive to be of general use. We need to be able to describe the taxonomic structure of the domain, to be able to model restrictions or constraints of the domain, and to be able to state and reason over a set of inference rules associated with the domain. We need to be able to describe an *ontology* of our domain.

The term ontology originated in the sphere of philosophy, where it signified the nature and the organisation of reality, ie. concerning the kinds of things that exist, and how to describe them. Our definition within Computer Science is more specific, and the most commonly cited definition has been provided to us by Tom Gruber in [6], where he defines an ontology as "an explicit and formal specification of a conceptualization". In other words, an ontology provides us with a shared understanding of a domain of interest. The fact that the specification is formal means that computers can perform reasoning about it. This in turn will improve the accuracy of searches, since a search engine can retrieve data regarding a precise concept, rather than a large collection of web pages based on keyword matching.

In relation to the Semantic Web, for us to share, reuse and reason about data we must provide a precise definition of the ontology, and represent it in a form that makes it amenable to machine processing. An ontology language should ideally extend existing standards such as XML and RDF/S, be of "adequate" expressive power, and provide efficient automated reasoning support. The most widely used ontology language is the "Web Ontology Language", which curiously has the acronym "OWL"[4]. Along with RDF/S, OWL is a W3C standard and augments RDFS with additional constraints such as localised domain and range constraints, cardinality and existence constraints, and transitive, inverse, and symmetric properties.

Adding a reasoning capability to an ontology language is tricky since there will be a trade-off between efficiency and expressiveness. Ultimately it depends on the nature and requirements of the end application, and it is for this reason that OWL offers three sublanguages,

- OWL Lite supports only a limited subset of OWL constructs and is computationally efficient,
- OWL DL is based on a first order logic called Description Logic,
- OWL Full offers the full compatibility with RDFS but at the price of computational tractability.

Examples of applications which could require very different levels of reasoning capabilities are described in the following section.

The top layers of the layer cake have received surprising little attention considering that they are crucial to successful deployment of Semantic Web applications. The proof layer involves the actual deductive process, representation of proofs, and proof validation. It allows applications to inquire why a particular conclusion has been reached, ie. they can give proof of their conclusions. The trust layer provides authentication of identity and evidence of the trustworthiness of data and services. It is supported through the use of digital signatures, recommendations by trusted agents, ratings by certification agencies etc.

C. Recent Trends and Technological Developments

As with any maturing technology, the architecture will not remain static. In 2006 Tim Berners Lee suggested an update to the layer cake diagram [2] which is shown in Figure 4, however this is just one of several proposed refinements. Some of the new features and languages which include the following.

Rules and Inferencing Systems. Alternative approaches to rule specification and inferencing are being developed. RIF (Rules Interchange Format) is a language for representing rules on the Web and for linking different rule-based systems together. The various formalisms are being extended in order to capture causal, probabilistic and temporal knowledge.

Database Support for RDF. As the volume of RDF data increases, it is necessary to provide the means to store, query and reason efficiently over the data. Database support for RDF and OWL is now available from Oracle (although at present the focus is on storage, rather than inferencing capabilities). Other open source products include 3Store[5] and Jena[6]. The specification of a query language for RDF, SPARQL, was adopted by the W3C in 2008.

RDF Extraction. The language GRDDL: ("Gleaning Resource Descriptions from Dialects of Languages") identifies when an XML document contains data compatible with RDF and provides transformations which can extract the data. Considering the volume of XML data available on the web, a means of converting this to RDF is clearly highly desirable.

[4] www.w3.org/2004/OWL

[5] http://sourceforge.net/projects/threestore/

[6] http://jena.sourceforge.net/

Ontology Language Developments. The OWL language was adapted as a standard in 2004. In 2007, work began on the definition of a new version, OWL 2 which includes easier query capabilities and efficient reasoning algorithms scaled to large datasets.

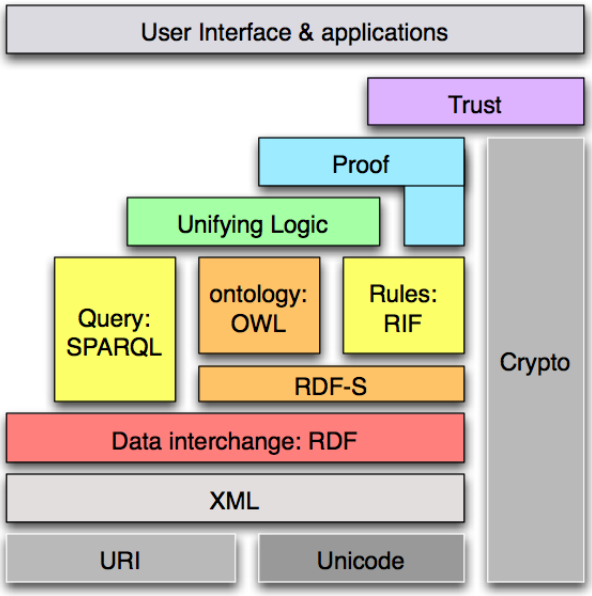


Figure 4. A Revised Semantic Web Layer Cake

IV. The Spectrum of Applications

Even though Semantic Web technology is in its infancy, there are a wide range of applications in existence. In this section we give a brief overview of some typical application areas.

E-Science Applications. Typically e-science describes scenarios involving large data collections requiring computationally intensive processing, and where the participants are distributed across the world. An infrastructure whereby scientists from different disciplines are able to share their insights and results is seen as critical, particularly when we consider the availability of large volumes of data becoming available online. The Gene Ontology[7] is a project aimed at standardizing the representation of genes across databases and species. Perhaps the most famous e-science project is the Human Genome Project[8] which identified the genes in human DNA and which includes over 500 datasets and tools. The International Virtual Observatory Alliance[9] makes available astronomical data from a number of digital archives.

Interoperation of Digital Libraries. Institutions such as libraries, universities, and museums have vast inventories of materials which are increasingly becoming available online. These systems are implemented using a range of different technologies, and although their aims are similar it is a huge challenge to enable the different institutions to access each other's catalogues. Ontologies are useful for providing shared descriptions of the objects, and ontology mapping techniques are being applied to achieve semantic interoperability [3].

Travel Information Systems. The goal of building an application which would allow a user to seamlessly book and plan the various elements of a trip (flights, hotel, car hire etc.) is highly desirable. Ontologies again could be used to arrive at a common understanding of terminology. The Open Travel Alliance is building XML based specifications which allow for the interchange of messages between companies. While this is a first step, an agreed ontology would be needed in order to achieve any meaningful interoperation.

Although many potential applications can be identified, there are less deployed at this time than we might expect. One possible reason is the lack of a common understanding of what the Semantic Web can offer, and more particularly what the role of ontology. At one end of the spectrum we find applications which take the "traditional", or AI view of inferencing, in which accuracy is paramount. Such applications arise in combinatorial chemistry for example, in which vast quantities of information on chemicals and their properties are analysed in order to identify useful new drugs. By coding the required drug's properties as assertions will reduce the number of samples which need to be constructed and manually analyzed by orders of magnitude. In cases such as these, the time taken to perform the inferencing is less important, since the trade-off will be a large reduction in the samples to be analyzed.

At the other end of the spectrum, we have "data centric" web applications which require a swift response to the user. Examples of this type of application include social network recommender systems such as Twine[10] which make use of ontologies to recommend their users to other individuals who may be of interest to them. While it is clear that a response must be generated for the user within a few seconds, we can observe too that there can be no logical proof of correctness and soundness of the answers generated in this type of case! Accordingly, the level of inferencing required in this type of application is minimal.

V. The Future: A Web of Data?

While we have stated that the Semantic Web focuses on data in contrast to the document centric view of the traditional web, this is not the complete picture. In order to realize value from putting data on the web, links need to be made in order to create a "web of data". Instead of having a web with pages that link to each other, we can have (with the same infrastructure) a data model with information on each entity distributed over the web.

The Linking Open Data [3] project aims to extend the collections of data being published on the web in RDF

[7] http://www.geneontology.org/index.shtml
[8] http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml
[9] www.ivoa.net
[10] www.twine.com

format and to create links between them. In a sense, this is analogous to traditional navigation between hypertext documents where the links are now the URIs contained in the RDF statements. Search engines could then query, rather than browse this information.

In a recent talk at the TDC 2009 conference[11], Tim Berners Lee gave a powerful motivation example for the project: scientists investigating the drug discovery for Alzheimer's disease needed to know which proteins were involved in signal transduction and were related to pyramidal neurons. Searching on Google returned 223,000 hits, but no document provided the answer as nobody had asked the question before. Posing the same question to the linked data produces 32 hits, each of which is a protein meeting the specified properties.

At the conception of the project in early 2007, there were a reported 200,000 RDF triples published. By May 2009 this had grown to 4.7 billion [dh]. Core datasets include

- DBpedia, a database extracted from Wikipedia containing over 274 million pieces of information. The knowledge base is constructed by analyzing the different types of structured information, such as the "infoboxes", tables, pictures etc.
- The DBLP Bibliography, which contains bibliographic information of academic papers,
- Geonames, which contains RDF descriptions of 6.5 million geographical features.

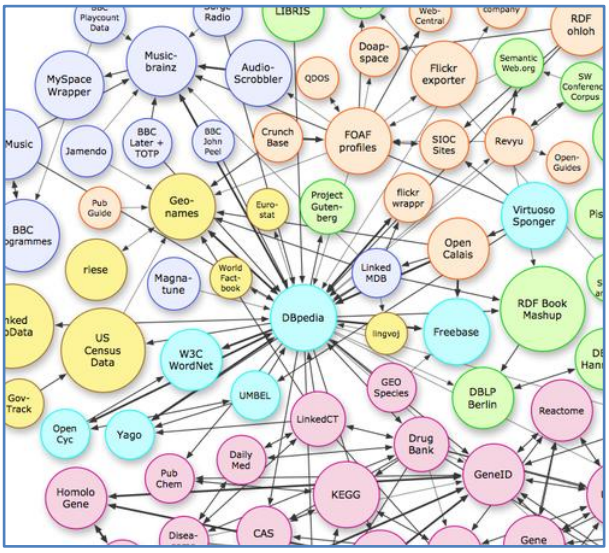


Figure 5. Web of Data (fragment)[12] [July 2009]

VI. Looking Ahead and Conclusions

So where is the Semantic Web? In a 2006 article [11], Tim Berners Lee agreed that the vision he described in the *Scientific American* article has not yet arrived. But perhaps it is arriving by stealth, under the guise of the "Web 3.0" umbrella. Confusion still abounds about the meaning of the term "Web 3.0", which has been variously described as being about the meaning of data, intelligent search, or a "personal assistant". This sounds like what the Semantic Web has to offer, but even if the terms do not become synonymous, it is clear that the Semantic Web will form a crucial component of Web 3.0 (or *vice versa*!).

The last five years have seen Semantic Web applications move from the research labs to the marketplace. While the use of ontologies has been flourishing in niche areas such as e-science for a number of years a recent survey by Hendler [7] shows a marked increase in the number of commercially focused semantic web products. The main industrial players are starting to take the technology more seriously. In August 2008, Microsoft bought Powerset, a semantic search engine, for a reported $100m.

As we have discussed, the "chicken and egg" dilemma is resolving itself with tens of billions of RDF triples now available on the web, and this number is continuing to increase exponentially.

Also, it is becoming easier for companies to enter the market of Semantic Web applications. There are now a wide range of open source applications such as Protégé[13] and Kowari[14] which provide building blocks for application development, making it more cost effective to develop Semantic Web products.

Some observers argue that the Semantic Web has failed to deliver its promise, arguing instead that the Web 2.0 genre of applications signifies the way forward. The Web 2.0 approach has made an enormous impact in recent years, but these applications could be developed and deployed more rapidly as their designers did not have the inconvenience of standards to adhere to. In this article we have demonstrated the steady infiltration from the research lab to the marketplace being made by the Semantic Web over the last decade. As the standards mature and the web of data expands, we are confident that the Semantic Web vision is set to become a reality.

References

[1] Berners-Lee T, Hendler J, Lassila O. 2001. The semantic web. In Scientific American, May 2001, available at http://www.scientificamerican.com/article.cfm?id=the-semantic-web

[2] Berners-Lee et al, 2006. A Framework for Web Science, Foundations and Trends in Web Science. Vol. 1, No 1.

[3] Chen H. 1999. Semantic Research for Digital Libraries, D-Lib Magazine, Vol. 5, No. 10, October 1999. http://www.dlib.org/dlib/october99/chen/10chen.html

[4] Davies J, Fensel D, van Harmelen F (eds). 2003. Towards the Semantic Web: Ontology-Driven Knowledge Management. John Wiley & Sons, Ltd.

[11] http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html
[12] http://en.wikipedia.org/wiki/File:Lod-datasets_2009-07-14_colored.png
[13] http://protege.stanford.edu/
[14] http://www.kowari.org/

[5] Fensel D, Hendler JA, Lieberman H, WahlsterW(eds). 2003. Spinning the Semantic Web: Bringing the World Wide Web to its Full Potential. MIT Press: Cambridge, MA. ISBN 0-262-06232-1.

[6] Gruber, T. 1993. Toward principles for the design of ontologies used for knowledge sharing. In Guarino N, Poli R (eds). International Workshop on Formal Ontology, Padova, Italy,

[7] Hendler, J., 2008. Linked Data: The Dark Side of the Semantic Web, (tutorial), 7th International Semantic Web Conference (ISWC08), Karlsruhe, Germany.

[8] Linking Open Data Wiki, available at http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

[9] Manning, C., Schütze, H., 1999. Foundations of statistical natural language processing. MIT Press.

[10] "Semantic Web - XML2000, slide 10". W3C. http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html.

[11] Shadbolt, N., Hall, W., Berners-Lee, T., 2006. The Semantic Web Revisited. IEEE Intelligent Systems. http://eprints.ecs.soton.ac.uk/12614/1/Semantic_Web_Revisted.pdf.

Exploiting Wikipedia as a Knowledge Base: Towards and Ontology of Movies

Rodrigo Alarcón, Octavio Sánchez, Víctor Mijangos

Grupo de Ingeniería Lingüística, Universidad Nacional Autónoma de México
Basamento de la Torre de Ingeniería, Ciudad Universitaria, México, D.F.
{ralarconm,osanchezv,vmijangosc}@iingen.unam.mx

Abstract. Wikipedia is a huge knowledge base growing every day due to the contribution of people all around the world. Some part of the information of each article is kept in a special, consistently and formatted table called *infobox*. In this article, we analyze the Wikipedia infoboxes of movies articles; we describe some of the problems that can make extracting information from these tables a difficult task. We also present a methodology to automatically extract information that could be useful towards the building of an ontology of movies from Wikipedia in Spanish.

Keywords: Wikipedia mining, Wikipedia infobox, ontology construction, semantic relation extraction, information extraction, natural language processing.

1 Introduction

Wikipedia is a free encyclopedia of open content that has become an important resource towards the construction of the Semantic Web. Since it beginnings, in the year 2001, the English version has achieve more than 2 million of articles, while the Spanish version has around 480 thousand of articles. All of the content has been written and edited by volunteers from different countries in many different languages, and it is covered by GFDL (GNU Free Document License), which makes possible to freely use them.

One important thing about the structure of Wikipedia is the social control executed by the community, which is able to avoid the spam, the nonsense and other kind of vandalism that is recurrent on some media sites. Besides, this same control makes possible to constantly increase the quality and precision of the articles.

Inside Wikipedia, there is an entry called *Wikipedia: Wikipedia in academic studies*[1], where it is possible to see the growth of academic interest in this encyclopedia. This interest is related to the use of Wikipedia on different academic studies and as a knowledge base for developing specific tools. On one hand, to mention a few, some works have focused on the social theme that represents Wikipedia [1] [2], some other have denounced inherent problems presented on this

[1] http://en.wikipedia.org/wiki/Academic_Research_on_Wikipedia.

kind of media sites [3], and others have obtained specific information and statistic data about the users [4]. On the other hand, Wikipedia has become a useful resource for the extraction of definitions, name entity recognition, machine translation or semantic relation extraction [5]. In this last field, Wikipedia represents a huge knowledge base that has made possible the developing of specific ontologies for the construction of the Semantic Web.

In this paper we present a work in process for the elaboration of an ontology of movies from Wikipedia on Spanish language. First we will briefly present an overview of some studies related to the use of Wikipedia for semantic relation extraction and ontology construction (2). Then we will explain the first step towards the elaboration of an ontology of movies (3). This step includes: a) the description of the so-called *infobox*, which is part of each movie of Wikipedia and contains specific data about the film (3.1); b) the specific relations to automatically extract (3.2); c) and our proposed XML schema to represent these relations (3.3). Finally, we will discuss our preliminary results and present the future work (4).

2 Wikipedia as a Semantic Knowledge Base

There is a growing interest of efforts to mine the information in Wikipedia for different purposes. As we have mentioned before, one of this interest is the extraction of semantic information that could be helpful on the process of *giving more meaning* to the Web. In Wikipedia, the meaning could be seen as the knowledge about things represented in different ways: definitions, descriptions, images, numeric data, etc. Furthermore, the meaning of each concept explained on the encyclopedia is related to the meaning of other concepts, which becomes a helpful semantic network to understand concepts on the field where they belong.

In this sense, Wikipedia represents a valuable source of knowledge to extract semantic information between concepts. A general overview of how Wikipedia could be used to extract concepts, relations, facts and descriptions can be found in [6]. Here, the authors explain the use of Wikipedia for natural language processing, information extraction and ontology building.

In [7], the authors describe a methodology that uses the links between categories to mine specific relations. They analyze some measures to infer relations and try to provide a semantic scheme in order to improve the search capabilities and to give the users meaningful suggestions to edit articles. In the same context, in [8] the authors use Wikipedia to develop a methodology for the automatic annotation of different semantic relations. This work is based on discovering lexical patterns that can be used to recognized specific relations between concepts. They evaluate the methodology by using a corpus and searching on it the relations founded in Wikipedia. Their results show that this kind of methodology could be a good starting point for automatic ontology construction.

The research presented in [9] shows how hyperlinked pages are used to generate a domain hierarchy by means of ranking articles that are strongly linked. These articles become a domain corpus for the automatic construction of an ontology. The same goal of obtaining ontologies through Wikipedia is described in [10], where the authors

apply machine learning techniques to improve the performance of a system that mines the *infoboxes*. Finally, in [11] we can found another example of the use of Wikipedia for ontology construction, specifically for document classification.

This is not, and does not pretend to be an extensive list of all the works made about semantic relation extraction or ontology construction from Wikipedia. Our main purpose is to state both the interest that has woken up in the area of extraction and organization of semantic information, and some of the automatic analyzes and procedures that are possible to develop taking into account Wikipedia's structure. Nevertheless, as we will see in this paper, this structure is often not well organized and makes it difficult to implement automatic processes.

3 Towards an Ontology of Movies

In order to develop an ontology of movies we have stated three main steps that can lead us to our purpose. The first one is to collect our input corpus from Wikipedia movies articles and the analysis of the infobox structure on them. After that, the second step is the delimitation and automatic extraction of specific semantic information. Finally, as a third step we consider the implementation of the extracted information into a XML schema that will conform the basis for another later annotation schema.

3.1 Movies infobox structure

The first step of our methodology was to conform a corpus from the articles of the *films by year* category. We use the *categories tree* option to find a list of the movies titles from the year 1892 to 2008[2]. After that, we use the *export pages* option to retrieve all the articles of this list. We found a total of 5,561 articles, where the opening and closing infoboxes tags ({{Fields…}}) was on 5,092 cases. This late number represents the total of articles from our corpus.

After that, we analyze the *infobox* of each entry. The *infobox* is a resource used on Wikipedia to summarize and group the information about specific data on some articles. In general words, its purpose is to make the information on a more available format and it can be use as a resource to other applications.

In Spanish language, there are 49 proposed fields for the infobox, where only two are consider as required: *film title* and *original title*. The infobox will be framed in {{Fields…}}, and each field inside will be preceded by a vertical bar "|" and followed by an equal sign "=" and the specific information. Fields without descriptions will remain empty after the equal sign. That means it will have the following structure:

```
| Field = description of the field
```

An example could be the next one:

```
| genre = Science fiction
```

[2] Data was collected on February 2009.

The whole fields used in the movies infoboxes from Wikipedia in Spanish can be found in table 1.

Table 1. Infobox template in Spanish.

Fields		
título original	diseño de producción	duración
título	guión	clasificación
índice	música	idioma
imagen	sonido	idioma2
nombre imagen	edición	idioma3
dirección	fotografía	idioma4
dirección2	montaje	productora
dirección3	vestuario	distribución
dirección4	efectos	presupuesto
dirección5	reparto	recaudación
dirección6	país	precedida_por
dirección7	país2	sucedida_por
dirección8	país3	imdb
dirección9	país4	filmaffinity
ayudantedirección	estreno	sincat
dirección artistica	estreno1	
producción	género	

From the table above we can see the different kind of information that the fields can introduce. We see information about *dirección* (direction), *estreno* (premiere), *idioma* (language, language2, language3, etc.), as well as *país* (country country2, country3, etc.), IMDb (Internet Movie Data Base) or Filmaffinity links (external Web sites with movies information).

The 49 fields from this table are the suggested ones in the official Wikipedia movies infobox template. Nevertheless, in our corpus we found several empty fields. We automatically found a total of 94,584 fields occurrences, while 30,742 cases where empty (32.48% of the whole occurrences).

Furthermore, one of the problems presented in the infoboxes is the lack of standardization. Some of the elements established by Wikipedia are written in an indistinctive way by the authors of the articles, while others have typographical errors. For example, the field *dirección* (direction) appears also as *director* (director); the field *título original* (original title) can be found as *título en España* (title in Spain), *título principal* (main title), *título traducido* (translated title), among others. More complicated is the case of *estreno* (premiere), which presents variations like *año* (year), *fecha* (date), *fecha de estreno* (premiere date), or *primera emisión* (first emission).

Typos are another common lack of standardization. For the field *género* (genre) we can find mistakes like **gènero*, **genero* or **genro*.

In the corpus we can also find the case of another fields that are not proposed in the original schema, such as *asistente de artes marciales* (martial arts assistant), *calificación* (qualification), *premios* (awards), *Myspace*, and so on. In this case we found a total of 205 non-official fields.

If we compare the schema in Spanish to the English one, we can notice that the latter infobox contains fewer fields, which probably allows to be more standardized at the moment to put it into practice. The fields of the movies infobox in English can be seen in table 2.

Table 2. Infobox template in English

Fields		
name	starring	country
image	music	language
image_size	cinematography	budget
caption	editing	gross
director	studio	preceded_by
producer	distributor	followed_by
writer	released	
narrator	runtime	

Here we can observe a total of 22 fields, comparing to the 49 in the Spanish template. It is important to notice the fact that most of another languages follow a similar structure like the one described for English. There is a similar template to the English movies infobox in French Wikipedia, with some added elements like format, awards, and *IMDb*. In Italian, the infobox determines general fields for different genres of films: *generic*, *animation* or *film a episodi* (films conformed from several short films), with specific fields for each genre; while in German, the fields specifies a more generic data, i.e., *title*, *original title*, *producer or cameraman*.

In infoboxes of different languages, the most common fields are *title*, *director* and *premiere*. There are also coincidences in other fields, for example *music* and *photography*. Between English and Spanish there is a coincidence in *preceded_by* and *followed_by*. Furthermore, in Spanish, as well as in French, there is the field of *IMDb*, while Italian or English do not include. However, in English links to *IMDb* or *Allmovie* can appear within the article as external links and not inside the template of the infobox. These external links are also a valuable information to extend the semantic data for an ontology, as they can add more information about the films that does not appears in Wikipedia, or be used to complete the empty fields of the infoboxes. Nevertheless, there is also no consistence between the occurrences of the tags with external links. In our corpus, the *IMDb* tag occurs approximately in the 80% of the articles, while *Filmaffinity* occurs around in the 5%.

3.2 Extracting specific relations data

Theoretically, the structure of the infoboxes contains information that should be exploited with relative easiness. We decide to automatically extract the *title*, *original title, director*, *premiere year* and *genre*, in order to generate a database with all of this information. Although, not all of this information is present on all the movies articles founded in the *films by year* category.

As we have mentioned before, there are some inconsistencies within the name of the fields, their completeness, or the way the authors write them. In the case of

director field, we found it with complete information in the 5,092 occurrences of the articles with infoboxes, however the field *genre* occurs only in 4,499 of these cases. Taking into account that the inconsistencies of the metadata make more difficult the process of automatic relation extraction from the films information, we achieve to obtain the data through the process we hereby describe.

From our corpus, we find out that 5,092 articles contained at least one director, although the field name from many of them was not the same and a review had to be made in order to compile a list of *ad hoc* synonyms for searching this specific field. The synset was formed by *dirección* (direction), *director* (director) and *dirigida* (directed). Also, after the equal sign that should follow the name of the field, the kind of following blanks was not always the same. Sometimes there were tabs; some other, more than one simple space; and even other, without spaces. Many of the director's names are also entries of Wikipedia, so many users decided to establish links to their names, using the symbol "[[" followed by the name of the director and closing with "]]". This has the purpose of specifying to the wikiengine that there is a link: [[link to the article]]. But not all of them had those brackets, and it caused troubles while parsing the data with the aim of recovering the director's name of the film associated with the title of the entry.

The same problems were founded when we tried to mine the original title of the movie. Despite the fact that this field does appear in all the infoboxes, not all of them appear with information, which means that there are articles with the *original title* field empty. It does not contain information in 195 articles occurrences in the corpus.

With the *premiere* field it was also problematic to extract the information, because most of the films had different words to express the premiere year, for example *año* (year), *fecha de estreno* (premiere date) or **añoacceso* (acces year). In this case we decide to mine only the *año* (year) and *estreno* (premiere) variants, because of the wide range of structural possibilities. We found that 23 films infoboxes do not contain a premiere year, sometimes it was in the title and sometimes were completely absent.

Other field we exploited was the one of *género* (genre), which also present some inconsistencies that could be attached to human errors at the time of transcribing the template. This field was empty on 593 occurrences in our corpus and is the more unused one.

Summarizing, we can find the number of occurrences for each field in table 3:

Table 3. Numerical data found over the analysis of infoboxes

Field name	Number of full fields	Number of empty fields
Director	5,092	0
Título	5,092	0
Título ID	5,092	0
Título orignal	4,897	195
Año	5,069	23
Género	4,499	593
Director	5,092	0
Título	5,092	0

From the table above we can see the three fields with empty information: *premiere* or *year*, *original title* and *genre*. The first one was empty only in 23 articles, while the

last one in more tan 500 cases. It is important to mention that the title of the movies was not obtained from the infobox but directly from the XML given by the Wikipedia, mainly because it is well demarcated by the labels *<title> </title>*; in the same way, we obtained the id used by the Wikipedia to identify each article.

Despite the inconsistencies and typos that make difficult the automatic process, in 4,499 cases all the information that we were trying to mine was complete. We consider that this number represents a good starting point to conform the basis of a first schema that could be later extended.

3.3 Proposed XML schema

With the data from the infoboxes that were exploited, we decided to generate a first XML scheme, which should give basic information about the film. This scheme can be expanded as we extend our extraction processes of the information contained in the Wikipedia articles.

To make this scheme, we decided to take *director* field as the root XML tag. The first tag will consist of the director's name. Taking into account that directors can have more than one film, we decided to introduce a *filmography* tag to include them. This last tag will include each *film* with *title*, *original title, year* and *genre* tags. On the opening *film* tag we added an attribute with Wikipedia's title id number. An example of the schema can be seen below.

Proposed XML schema for the organization of movies data on Spanish Wikipedia.

```
<director>
    <name>Brian de Palma</name>
    <filmography>
        <film wiki_id="2022905">
            <title>Carrie: la ira</title>
            <original_title>The Rage: Carrie 2 </original_title>
            <year>1999</year>
            <genre>Terror </genre>
        </film>
        <film wiki_id="587196">
            <title>La Dalia Negra</title>
            <original_title>The Black Dahlia</original_title>
            <year>2006</year>
            <genre>Crimen, Misterio, Thriller</genre>
        </film>
        …
        …
        …
    </filmography>
</director>
```

As we can see in this example, the root tag is *<director></director>*. It is followed by the director's name tag *<name></ name >*. At the same level there is the

tag *<filmography></ filmography>*. This tag nests the film tag *<film wiki_id=""></film>*, which contains the relevant information of each film: *<title></title>*, *<original_title></original_title>*, *<year></year>* and *<genre></genre>*.

Based on the XML scheme, relational databases can be generated to manipulate the information that we have considered at this first stage of the ontology construction. As we have said, this is not the full final scheme because as more data is extracted, the more can be added. This scheme is currently based on Wikipedia films articles in Spanish language, however it can be extended to fit another kind of relevant information, for example the country, external links (*IMDb*) or the id of directors or genre from Wikipedia. Furthermore, it will be possible to use this scheme in order to exploit Wikipedia in other languages, which could make possible to fill the empty fields in one language by relating them with the information on another languages, as well as to make multilingual queries.

4 Conclusions and future work

Nowadays, Wikipedia can be explored with the aim of obtaining information on different ways. The information added in a manual way by the users is generally well organized and semi-structured. Also, many entries from Wikipedia have infoboxes with summarized specific information about the theme treated in the article. We have mentioned that the structure of Wikipedia has made possible to exploit the information in order to extract semantic data. The extraction of semantic relations is one of the growing interests aiming to the construction of the Semantic Web.

Even so the structure of Wikipedia, we have noticed some specific problems on automatically exploiting it. To summarize a few, there are: a) the fact that the field's names are not respected; b) typos by human errors; c) lack of information; and d) differences on the infobox structure between languages. The latter should not be seen as a problem, however it would be advantageous to have standard fields on different languages.

Aiming to the standardization idea, it would be useful that the Wikipedia's process of writing or editing an article use a check-bot to confirm the information of the infoboxes templates. Thus, the fields not belonging to the template would be alerted, as well as typos on the field names. Furthermore, the same check-bot could be used to seek the existing fields looking for inconsistencies in the infoboxes or the whole articles.

The work that we have presented here is a first approach towards the elaboration of an ontology of movies from the Wikipedia in Spanish. We have showed the kind of semantic relations that are possible to mine, as well as a first scheme to represent them. We are conscious that this scheme may well be improved for achieving a complete ontology of movies. The future work will include: a) to define a scheme to represent subject, relation and predicates between the extracted information, for example a RDF scheme; b) to implement this new scheme for making the information available and sharing it with systems dedicated to the construction of the Semantic

Web; c) to develop a movie-ontology query system capable of retrieve the information on specific ways related to *directors*, *titles*, *genres* and *years* fields.

Acknowledgments

This research was made possible by the financial support of CONACYT (82050) and DGAPA-PAPIIT (IN403108). The authors wish to thank Sarahi Abrego Romero for the proofreading of this paper.

References

1. Kittur, A., Chi, E., Pendleton, B. A., Suh, B., Mytkowicz, T.: Power of the Few vs. Wisdom of the Crowd: Wikipedia and the Rise of the Bourgeoisie. In: 25th Annual ACM Conference on Human Factors in Computing Systems (CHI 2007), ACM, New York (2007)
2. Suh, B., Chi, E. H., Pendleton, B. A., & Kittur, A.: Us vs. Them: Understanding Social Dynamics in Wikipedia with Revert Graph Visualizations. In: Visual Analytics Science and Technology. pp. 163-170, IEEE-Press, New York (2007)
3. Potthast, M.,Stein, B., Anderka, M.: Automatic Vandalism Detection in Wikipedia. In: 30th European Conference on IR Research, ECIR 2008, pp. 663-668, Glasgow (2008)
4. Stein, K., Hess, C.: Does it matter who contributes: a study on featured articles in the german wikipedia. In: Proceedings of the 18th conference on Hypertext and hypermedia, pp. 171-174, ACM, New York (2007).
5. 99 Wikipedia Sources Aiding the Semantic Web. AI3, http://www.mkbergman.com/?p=417
6. Medelyan, O., Milne, D., Legg, C., Witten, I.A.: Mining meaning from Wikipedia. Hamilton, (2008)
7. Chernov, S., Iofciu, T., Nejdl, W., Zhou, X.: Extracting Semantic Relationships between Wikipedia Categories. In: Proceedings of the 1st Workshop on Semantic Wikis - From Wiki to Semantics, ESWC2006, Budva (2006)
8. Ruiz-casado, M., Alfonseca, E., Castells, P.: From Wikipedia to Semantic Relationships: a Semi-automated Annotation Approach. In: Proceedings of the 1st Workshop on Semantic Wikis - From Wiki to Semantics, ESWC2006, Budva (2006)
9. Cui, G., Lu, Q., Li, W., Chen, Y.: Corpus Exploitation from Wikipedia for Ontology Construction. Conference on Language Resources and Evaluation, LREC2008, Morocco (2008)
10. Wu, F., Weld, D. S.: Automatically Refining the Wikipedia Infobox Ontology. In 17th International World Wide Web Conference, Beijing (2008)
11. Kozlova, N.: Automatic Ontology Extraction for Document Classification. Ma. Thesis, Saarland University (2006)

Translation of Verbal Expressions and Context of Use Extraction Through a Corpus on Web

Arturo Velasco, María J. Somodevilla, and Ivo. H. Pineda

Facultad de Ciencias de la Computación, Universidad Autónoma de Puebla
arturoezak@hotmail.com, {mariasg, ipineda}@cs.buap.mx

Abstract. The group of fixed expressions constitutes an important part of the lexical system. This group is defined as the stable combination of two or more elements that is not possible to establish a meaning from its constituents, in addition to a grammatical structure which can move away from the language rules. In this paper a method of processing, translation and context of use extraction of verbal expressions (subsets of the fixed expressions) into a diatopic system of the Spanish language is presented. The architecture proposed is organized in three modules: the database, whose objective is to be able to store essential characteristics of them; the corpus, that contains digital texts and transcribed oral language; and the extraction expressions module, which extract examples on the corpus.

1. Introduction

The *UFs* or *fixed expressions* are expressions consisting of two or more words whose meaning can not be inferred from the union of the significance of each of the lexical elements that constitute it. Zuluaga A., describes two basic characteristics that have fixed expressions: idiomaticity, characteristics that are peculiar and unique to a specific language or sub-language, including some socio-cultural traits; and fixation, the property that has the expressions of being reproduced in the speech like previously defined combinations, i.e. they present certain order in their syntactic structure [1].

Other important characteristics that have the *UFs* are: high frequency of report of their constituent elements, absence of grammatical rules in the expressions and translation problems.

On the other hand, due to the lack of agreement between linguists to establish limits of research of the phraseology and terminology used in this area, we decided to follow Alberto Zuluaga's work [1].

Zuluaga carries out a classification of the *UFs* based on the actions of the expressions in the speech. In the first group, Zuluaga establishes the locutions like a stable combination of two or more terms that work as an element in sentences to level of lexeme or syntagm. Inside this classification (locutions) he separates those that are in use as grammatical instruments and the expressions that possess semantic sense

(lexical units). The subset of the *UFs* object of study in this work are the *locutions* and *verbal syntagms* whom belong to the units with lexical sense. The verbal locution is equivalent to lexemes, e.g.: *pasar a mejor vida* (to die) or *echar una mano* (to help) and the verbal syntagm are equivalent to syntagms e.g.: *pagar los platos rotos* (to suffer the consequences of something).

Considering those problems mentioned above and the *UFs* taxonomy of Zuluaga, it's proposed to develop a *Diatopic Verbal Expressions Digital Dictionary* (DIVEDD) for Spanish Language (diatopic subsystems of Spain and México) in order to enable the process of translation of verbal expressions (verbal locutions and verbal syntagms) in both subsystems. This prototype uses regular expressions and keywords, generating synonyms and variants expressions, finally, shows through a Corpus, examples of real use.

The paper is organized as follows. The second section describes related work with processing and translation of *UFs*; third section presents the architecture of the DIVEDD; the results are showed in the fourth section; and finally, conclusions are presented in the fifth section.

2. Related Work

The group of *UFs* constitutes an important part of the lexical system, where monolingual and bilingual dictionaries only capture certain number of units, often reduced, to an alphabetical process of selection and random description [2]. In México there are not recent works of compilation of expressions, some of them are: the *Diccionario breve de mexicanismos* [3], the *Diccionario ejemplificado de mexicanismos* [4], and the *Diccionario del español usual en México* [5]. The lack of strict rules at the time of integrate these dictionaries brought the introduction of different subsets of *UFs*.

There are some works related with translation of expressions in the Spanish language such as *Recopilación de proverbios*, proverbs which were translated into four languages (English, French, German and Italian) [6]. *In Spagnolo-Italiano: Espressioni idiomatiche e proverbi*, there are a summary of idiomatic expressions, proverbs and Spanish and Italian pragmatic sentences [7]. In [8] there are *877 refranes españoles*, sayings with their correspondence Catalan, Galician, Basque, English and French. Finally, *Divergencias en la traducción de expresiones idiomáticas y refranes* by Corpas Pastor [9], that provides a more systematic methodology for the translation of expressions between French and Spanish (Spain). This model of bibliographical record considers different uses, the level of the speaker's registration, antonyms, synonyms, source of the expression and examples among other data. This work was considered as a starting point and taking the benefits of a corpus for showing use of actual situations of expressions.

3. Prototype Architecture

The architecture proposed of the DIVEDD is organized in three modules: the database, that contains the essential characteristics of the expression; the corpus, that contains in this first stage of digital texts and transcribed oral language; and the expansion expressions module, that is complemented with a list of stop-words and a database storing verbal conjugations. In the fig. 1the architecture of the DIVEDD is shown.

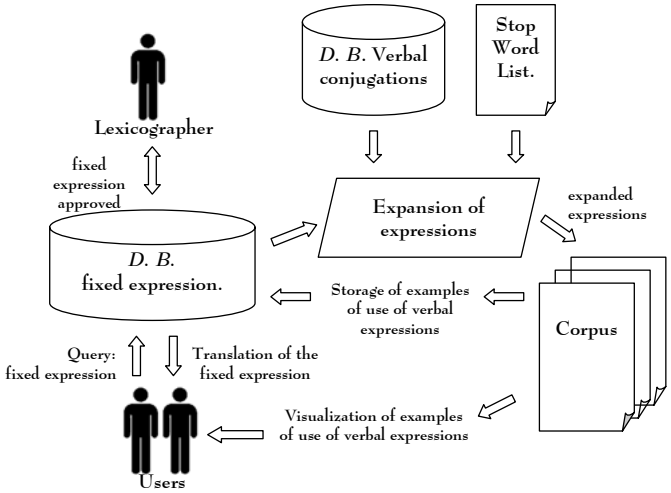


Fig. 1. DIVEDD architecture.

3.1 Variants, Synonyms and +Frequents Expressions

Variants: those expressions that vary or omit any of its closed lexical elements without having semantic change.
Synonyms: those expressions that have changed in their non closed lexical element i.e. key-word or those which do not contain any element in common, but they do not have a semantic change.
+Frequent: the most used or most likely expression to appear in the dictionary, within a set of variations. Thus, +Frecuent expression is taken as representative.

Table 1 shows an example of synonym expressions, therefore, the expressions: *ir al bote*, *ir al tambo*, *ir a la sombra* and *ir tras las rajas* are synonyms, because their keywords (*kw*) changes but they have the same definition. The same situation applies to the expressions *hacer la barba*, *hacer la pelota* and *hacer la rosca*, but in addition, *hacer la barba* is the translation into Spanish (México) of the expressions of Spain *hacer la pelota* and *hacer la rosca*.

Table 1. Handling synonym expressions in the DIVEDD.

	+Frequent_ Verbal_ Expression	*Definition*	*Key-word*	*Thematic_ Field*	*Linguistic_ Record*	*Country*
Synonyms	Ir al bote	Meter a alguien en la cárcel	Bote	Behavior	Informal	México
	Ir al tambo	Meter a alguien en la cárcel	Tambo	Behavior	Informal	México
	Ir a la sombra	Meter a alguien en la cárcel	Sombra	Behavior	Informal	México
	Ir tras las rejas	Meter a alguien en la cárcel	Rejas	Behavior	Informal	México
	Hacer la barba	Lisonjear a alguien	Barba	Behavior	Informal	México
	Hacer la pelota	Lisonjear a alguien	Pelota	Behavior	Informal	Spain
	Hacer la rosca	Lisonjear a alguien	Rosca	Behavior	Informal	Spain

Table 2 shows variants through regular expressions. A regular expression is a set of pattern matching rules encoded in a string according to certain syntax rules [10]. Thus, it is possible to describe or represent a set of strings without need to enumerate all of its elements. The operators used in the right column of Table 2 are described in table 4 in section 3.4, *Generation of Synonyms and Variants*.

Table 2. Variants expressions in the database of the DIVEDD.

+Frequent_ Verbal_ Expression	*Definition*	*Key-word*	*Variants_Verbal_Expression*
Ir al bote	Meter a alguien en la cárcel	Bote	[ir,llevar,meter] (al) {bote} [refundir] (en_el) {bote}
Ir al tambo	Meter a alguien en la cárcel	Tambo	[ir,llevar,meter] (al) {tambo} [refundir] (en_el) {tambo}
Ir a la sombra	Meter a alguien en la cárcel	Sombra	[ir,llevar,meter] (a_la) {sombra} [refundir] (en_la) {sombra}
Ir tras las rejas	Meter a alguien en la cárcel	Rejas	[ir,llevar,meter,refundir] (tras_las) {rejas}

3.2 The Database

This module is based on a relational model that provides mechanisms that guarantee to avoid duplicity of records and inconsistency problems; it also guarantees the referential integrity and favors improvements of processing of the expressions. Table 3 shows the most important attributes of verbal expressions to store.

Table 3. More important attributes of the verbal expressions.

ATTRIBUTE	DESCRPTION
Verb	Main verb used in the expression.
Canonical_ Verbal_ Expression	*Locution* or *verbal syntagm* in its canonical form.
Definition	Definition of the verbal expression. Field used to make the translation among the diatopic verbal expressions.
Source	Resource where the expression was extracted.
Use_ Frecuency	The number of frequencies of appearance of the expression in the corpus.
Linguistic_ Record	Level of registration of the expression.
Country	Country of origin of the expression.
Region	area or region of use of the expression.
Thematic_ Field	Thematic field of the expression
Key-Word	Alexical component unit of the expression. Useful to distinguish among synonym expressions.
Variant_ Verbal_ Expression	Field of the table *Variant* that stores the variants of the canonical verbal expressions.
Example	Field of the table *Examples* that stores the examples provided by the lexicographer and extracted of the corpus.

3.3 Corpus Processing for DIVEDD

The corpus of the DIVEDD is conformed by written language and transcribed oral language [11], based on the recommendations of Sinclair J. [12], [13]. The subcorpus of written language is built from digital Mexican newspapers (four sections: *local news*, *police section*, *opinion section* and *shows section* over geographical limitation in the states of D.F., Mexico, Hidalgo, Morelos, Puebla and Tlaxcala). The subcorpus of transcribed oral language is conformed by the Sociolinguistic Corpus of the México City (CSCM) [14]. Fig. 2 shows the processing for the last one subcorpus.

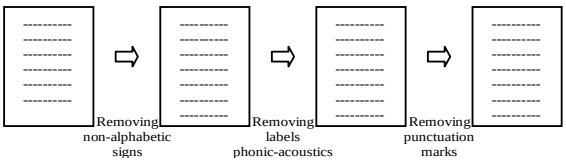


Fig. 2. Corpus processing.

3.4 Extraction Expressions Module

The module expressions expansion serves as a liaison between the DB and the corpus of DIDEVD. This module has two contributions: generate synonyms and variations of the +Frequent expression stored in the DB; and extract examples of actual usage throw the corpus.

Generation of Synonyms and Variants. The generation of synonyms and variants of a *+frequent* expression requires the entry of the possible combinations that can occur between *kw´s* and *connectors-words* or *stop-words* (*sw*). To describe all these expressions without the need to enumerate each one of them, the regular expressions are used. The operators are shown in Table 4.

Table 4. Operators of the regular expressions.

Operator	Function
'[' y ']'	Denotes the set of verbs that can be used in the expression.
'(' y ')'	Denotes the set of connector–words between the verb and key-word.
'{' y '}'	Denotes the set of key-words that are used in the expression.
','	Separator of a set of words (verbs, key-words, connector–words). Can be use ',' instead of '|'.
'|'	Performs the same function as ','.
'_'	Joint two or more nonseparable words in an expression
' '	The blank space denotes the separation between groups.

Considering the operators used in regular expressions specified in Table 4, the canonical verbal expression formed by *hablar más que un loro*, where the *kw* is *loro* and the regular expressions are denoted by:

[hablar,platicar] (más_que_un,como,como_un) {loro,perico,merolico}
[hablar,platicar] (más_que_una,como_una) {cotorra}

The set of variants of hablar más que un loro are: *hablar como loro, hablar como un loro, platicar más que un loro, platicar como loro* and *platicar como un loro.*

The set of all its synonyms are: *hablar más que un perico, hablar como perico, hablar como un perico, platicar más que un merolico, platicar como merolico, platicar como un merolico, hablar más que una cotorra, hablar como una cotorra, platicar más que una cotorra* and *platicar como una cotorra*.

As it can be seen the properties of regular expressions help us to match any possible variation of the expressions in the corpus without necessity of having enumerated each one.

Extraction of Usage Examples. The second contribution of extraction expressions module is the search for examples of real use of expressions stored in the DB.
The search can be performed by matching between the expression and a fragment of the corpus. The second way is to find expression through similarity. Based on the premise that the *kw* is crucial in the expression,

The first step is to identify all the words that have a high degree of similarity with the *kw* was carried out. The similarity function between two strings is described in [15] and showed below.

The second step is to identify words that preceding to the *kw*. Only *verbs* and *sw´s* are accepted. Another word unidentified will provoke that the fragment is rejected. The process consists of a retreat from the position of the *kw*. Ends in a satisfactory manner when encountering a *verb* and *sw`s* or if two verbs (an auxiliary and non auxiliary) and sw´s are matched.

```
//Similarity of two strings; return the percentage
function similarity($s1, $s2)
{
  $m = strlen($s1);
  $n = strlen($s2);
  $matrix = array(array($m),array($n));
  for($i=1; $i < $m; $i++) $matrix[$i][0]=0;
  for($j=0; $j < $n; $j++) $matrix[0][$j]=0;
  for ($i=1; $i <= $m; $i++) {
   for ($j=1; $j <= $n; $j++) {
     if ($s1[$i-1]==$s2[$j-1])
       $matrix[$i][$j] = $matrix[$i-1][$j-1] + 1;
     else if ($matrix[$i-1][$j] >= $matrix[$i][$j-1])
       $matrix[$i][$j] = $matrix[$i-1][$j];
     else
       $matrix[$i][$j] = $matrix[$i][$j-1];
   }
  }
  $avgs = ($m + $n) / 2;
  return ($matrix[$m][$n] / $avgs) * 100;
}
```

4. Results

The extraction process using regular expressions have shown little recovery since it requires a tie with the exact expressions given by the lexicographer. On another hand, the extraction process by similarity functions between *kw´s* and words in the corpus heralds not only an examples extraction process; also, variant expressions can be extracted to perform in a future work, the reverse process, i.e. creating regular expressions.

Thus, besides of the similarity applied and described in [15], the Levenshtein similarity was applied. This function calculates the minimum number of operations (insertion, deletion or substitution) required to transform one string into another. The results are not as regulars as [15], and has trouble distinguishing different *kw´s*, and grouping *kw´s* of which varies only in gender and number.

The DIVEDD database was developed in MySQL 5.0.18. All the processing is implemented in PHP 5.1.2. The fig. 3 shows the DIVEDD interface.

Verbo	Expresión Verbal	Definición	P. Clave	Camp. Tem.	Niv. Reg.	País
aventar	aventar a alguien flores	adular, lisonjear	flores	comportamiento	culto	México
caer	caer bien a alguien	obtener buena acogida	bien	comunicación	culto	México
chingar	chingar algo o a alguien	descomponer algo, importunar, molestar	chingar	comportamiento	informal	México
dar	dar a alguien el avión	no prestar atención	avión	comportamiento	culto	México
dar	dar chance	dar permiso, oportunidad	chance	comunicación	informal	México
dar	dar la cara	salir a su defensa	madre	comportamiento	estándar	México
dar	dar lana a alguien	dar dinero	lana	comunicación	informal	México
dar	dar un rol	paseo, viaje	rol	comunicación	estándar	México
dejar	dejar algo botado	demeritarle importancia	botado	comportamiento	estándar	México
dejar	dejar algo o alguien por la paz	no inquietarle ni molestarle, dejar en paz a alguien	paz	comportamiento	culto	México
dejar	dejar plantado a alguien	hacer esperar a alguien sin acudir a la cita	plantado	comportamiento	estándar	México
echar	echar la culpa	atribuirle la falta o delito que se presume ha cometido	culpa	comportamiento	estándar	México
estar	estar algo cabrón	difícil, complicado	cabrón	descripción	informal	México
estar	estar algo canijo	difícil, Mala persona	canijo	descripción	estándar	México
estar	estar algo cañon	muy bien, estupendo	cañon	descripción	informal	México
estar	estar algo chido	bueno, muy bueno	chido	descripción	estándar	México
estar	estar algo chingón	bueno, muy bueno	chingón	descripción	informal	México
estar	estar algo en chino	difícil de entender	chino	descripción	culto	España México
estar	estar algo muy colgado	lejano en distancia, tiempo	colgado	descripción	estándar	México
estar	estar algo muy pesado	violento, insufrible, difícil de soportar	pesado	descripción	estándar	México

Fig. 3. DIVEDD interface.

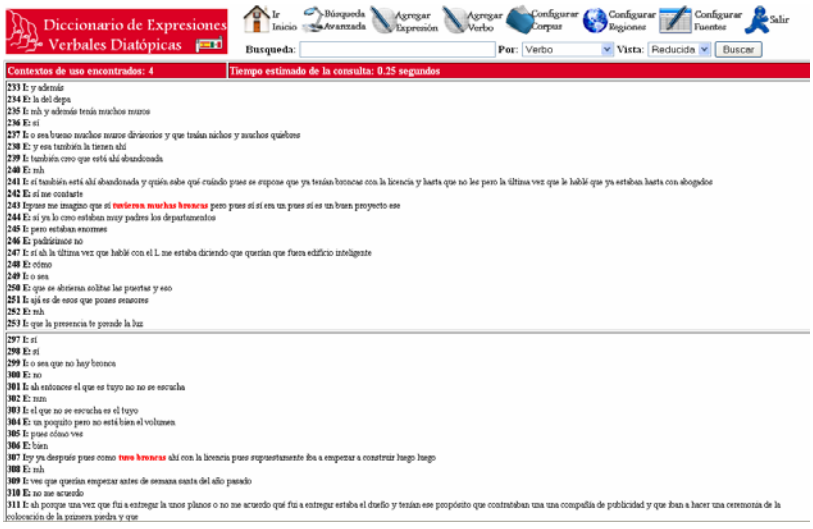


Fig. 4. Extraction of real examples of *tener broncas*.

5. Conclusions and ongoing work

The DIVEDD appears to be a system for human translation assisted by computer, providing a definition and basic characteristics of verbal expressions. The DIVEDD does not try to be a detailed dictionary but it is as a mechanism reliable of storage of

phraseological information enforcing structure and integrity of data, reducing times of search and translation. On the other side, the mechanisms of search expressions by expressions´ attributes and its combinations make the DIVEDD a flexible tool.

Finally, note that the extraction process starting from similarity on *kw´s* showed more encouraging results, but with *noise* (information not relevant to the query), because there are parts of texts recovered, that are not relevance to the phraseology, i.e. there are non-verbal expressions. We will work on linguistic heuristics to reduce the noise.

References

1. Zuluaga A. Introducción al estudio de las expresiones fijas. Frankfurt: Peter Lang. (1980)
2. Mogorrón H. Los diccionarios electrónicos fraseológicos, perspectivas para la lengua y la traducción. Universidad de Alicante (2004)
3. Gómez de Silva G. Diccionario breve de Mexicanismos. 1a ed., México, FCE. (2001)
4. Steel B. Breve Diccionario Ejemplificado de Mexicanismos (2000)
5. Lara L. Diccionario del español usual en México. 1a ed. ISBN: 9789681207045 (2003)
6. Casado, M. L., Agueda, S., Agueda, B. and Corral, J. Recopilación de proverbios. Alcobendas. ISBN: 9788471436450 (1998)
7. Zamora M. Spagnolo-italiano: espressioni idiomatiche e proverbi, Milano, EGEA (1997)
8. Sevilla M. and Cantera J. 877 refranes españoles con su correspondencia catalana, gallega, vasca, francesa e inglesa. Madrid: EUNSA (1998)
9. Sevilla M. Divergencias en la traducción de expresiones idiomáticas y refranes (francés-español) (1999)
10. Stubblebine T. Regular Expression, Pocket referente. O`really 2nd edition (2007)
11. Procházková P. Fundamentos de la lingüística de corpus, concepción de los corpus y métodos de investigación con corpus (2006).
12. Sinclair J. Preliminary Recommendations on Corpus Typology. EAGLES Document EAG-TCWG-CTYP/P. (1996)
13. Sinclair J. Developing Linguistic Corpora: a Guide to Good Practice Corpus and Text—Basic Principles (2004)
14. Colegio de México, http://lef.colmex.mx/Sociolinguistica/CSCM/Corpus.htm
15. Oliver J. DecisionGraphs-An Extension of Decision Tres. TechnicalReport No:92 /173 (1993)

Dynamic Concept-Based Taxonomy used for image recovery based on their textual description

Jaime Lara, María de la Concepción Pérez de Celis, David Pinto

Facultad de Ciencias de la Computación,
Benemérita Universidad Autónoma de Puebla,
Puebla, México.
{jlara, cperezdecelis, dpinto}@cs.buap.mx

Abstract. In this paper, we will describe a methodology for the development of a search system based on the usage of textual descriptions in order to recover images using a dynamic taxonomy. This taxonomy is stored in a relational database that is part of a system that allows the user to explore and refine his search, using a navigation tree. We propose the usage of information recovery techniques to extract a controlled vocabulary and defining concepts that will be structured in hierarchies with the aid of a thesaurus, in order to automatically generate facets. Such facets will then be linked with the studied objects using indexes. We have applied this model on a collection of artworks, linking an image with a textual description in the Spanish language. The experimental results show the advantages of such a model.

Keywords: automatic generation of taxonomies, dynamic taxonomies, faceted classification, faceted taxonomies, information recovery, navigation trees.

1 Introduction

An alternative to classical search engines (keyword-based search) is to allow the user to navigate through contents. This process of navigation requires the user to know the way in which the information is organized. A practical solution to this inconvenient is the use of a pre-established taxonomy. We must point out that taxonomies require the organization of contents around a certain knowledge domain. From our point of view, a taxonomy improves the recovery of information on specific topics, but it is still a rigid solution since it does not allow for the possibility of restructuring the users searches as he advances in the recovery of results. As such, the user cannot connect the recovered objects with other knowledge domains within such objects may be connected.

A possible solution that allows the user to be guided and also by self may interact and refine his search is the usage of what is known as a faceted taxonomy. Such taxonomy allows the information to be structured and accessed in more than one dimension. This benefits the user because he can easily and intuitively localize and explore the information using the different approaches provided by the taxonomy. In this work we will focus on the methodology used to build a taxonomy which will be

the base for the implementation of a system of faceted classification, which will then allow the user to build his own knowledge map. In the following sections we will present some of the work linked to the design and construction of taxonomies and we will analyze its usage in the recovery of information. Later on, we will discuss the methodology used for the confirmation of the taxonomy, with emphasis on the recovery of objects based on their textual descriptions. We will present the results we have obtained and we will emphasize the importance of conceptualizing the operations that allow for the creation of open taxonomies.

2 Related Works

An information retrieval system in which one has a collection of objects that have diverse properties, the relevance of such properties varies depending on the user. As a matter of fact, a solution for finding information regarding a particular topic consists in asking an expert on the subject. Due to the fact that, generally, this is not possible, one has to recur to the use of taxonomies.

The importance of taxonomies lies in the possibility of them being used as a triplet (classification scheme, semantic interpretation, knowledge map [1]) in the organization and recovery of collected objects (possibly in a database).

Taxonomies can be organized under different structures: lists, hierarchies, poli-hierarchies, multi-dimensional matrix and facets, among others. Among them, poli-hierarchies, multi-dimensional matrix and facets are linked to the possibilities of objects belonging to more than one category.

Characterizing objects under diverse categories or characteristics allows us to create a metadata model were the objects and their characteristics take on a new meaning. It is because of such a phenomenon that faceted taxonomies can take advantage of the way in which metadata behave. Metadata models are a collection of information built on some type of object, or on a part of such an object. For example, the name (or title) of an artwork, its author, date, image, textual description (denotation), interpretation (connotation) or genre, represent metadata that can be associated to a cultural object.

As such, every element in the metadata scheme can be incorporated as a concept of the faceted taxonomy, and thus it can be recovered using a search engine. Therefore it is possible to access an object under any of the dimensions under which it has been classified. If we consider the previous example, we could recover a cultural object by its genre, connotation or denotation or we could navigate through the different facets, taking into account that they are orthogonal among themselves.

Sacco [2], [3] introduced the concept of dynamic taxonomies and the notion that it could withstand the incorporation of facets that, in themselves, require an independent taxonomy for their description. Due to the fact that our objective is the recovery of images based on their textual description, we will use the fundamentals of dynamic taxonomies but we will also extend the domain of the data modeling created by Sacco, because we will include textual objects.

In the literature we can find two interesting approaches for the management of textual objects. The first such approach assumes the existence of various topics over

which one wishes to generate taxonomy, and because of that the algorithms are designed to extract the terms and concepts linked to such topics from the documents. The MindMap system [4], in particular, proposes the generation of multiple taxonomies for any collection of documents, each one with unique topics. This multiple taxonomies are visualized in the system as an integrated tool, thus obtaining a system that allows the organization of information in multiple ways. In this system, the classification of a document under any taxonomy depends on its similarities with other documents. For this purpose, a system of spatial coordinates is used, in which the similarities are determined by the proximity between the coordinates of each document. As we have pointed out before, in this approach each one of the multiple taxonomies requires a series of keywords, associated to the concepts under which the classification of the documents will be structured, in order for the analysis to begin.

In contrast, the second approach is centered in determining the different facets and defining their taxonomy from the textual analysis of the collected documents by using text analysis. A good example is the algorithm used by the Flamenco project [5] of Berkley University for the automatic generation of facets using WordNet over a corpus written in English.

3 Our approach

During the development of the information system for the management of cultural objects, we have implemented the metadata model proposed by CCO [6]. However, when extending this model with the metadata of genre, connotation and denotation, the inclusion of a textual description of the objects was necessary. This fact determines the necessity of a methodology that allows for the generation of a taxonomic structure of the facets, under which the different terms included in the description of the objects will be classified. We must point out that such a methodology can be extended to any corpus of descriptive documents over which one means to develop a taxonomy. Our proposal consists in using techniques and algorithms employed in information retrieval to generate a faceted classification, which allows for the association of part of a document to the different facets. Such facets will be generated from a thesaurus linked with a “controlled vocabulary”. This vocabulary, in turn, will be extracted fro the processing of different texts from the target collection. This process is described in detail in the following sections.

3.1 Approach by dynamic taxonomy

The largest difference between a conventional taxonomy and a dynamic taxonomy is that the former are monodimensional (one element is classified under one concept), while the former are multidimensional (Fig. 1). The term *dynamic* reflects the ability of the taxonomy to adapt to different approaches, perspectives and interests.

Sacco [2], [3] established the following inference rule: *Two concepts A and B are related iff there is at least one item D in the infobase which is classified at the same time under A (or under one of A’s descendants) and under B (or under one of B’s descendants).*

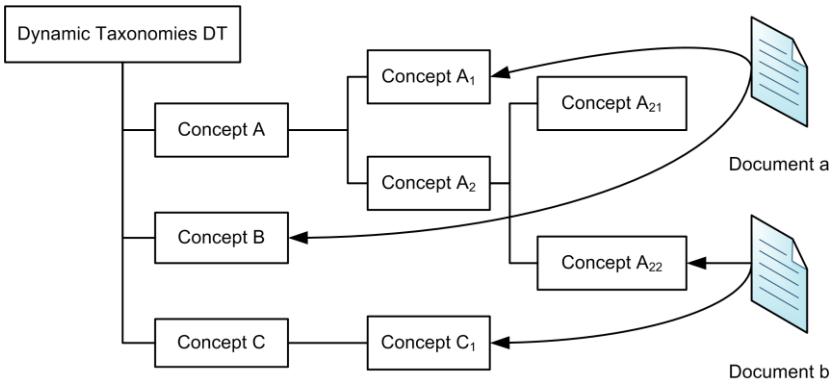


Fig. 1. Multidimensionality in dynamic taxonomies.

The left side of Fig. 2 shows a set of data classified under two facets, and in the right side can be appreciated the change that the taxonomies of our facets go through when we focus our search on the concepts B and H.

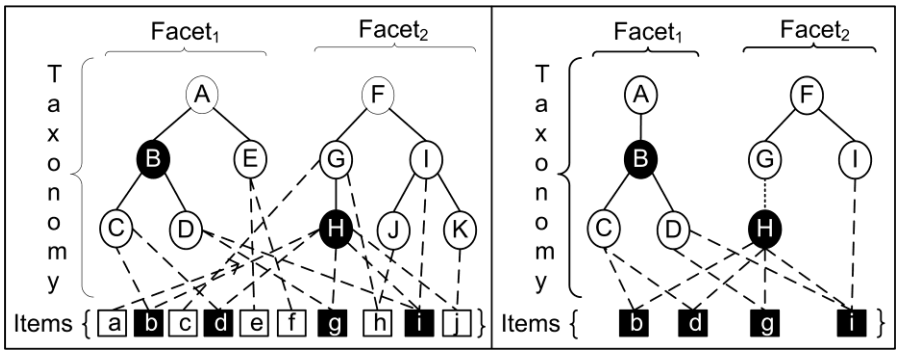
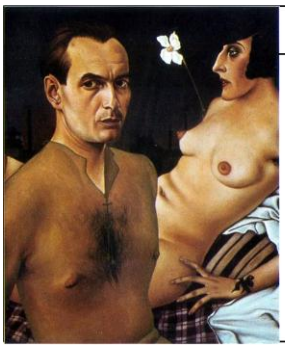


Fig. 2. Left side: Data set {a,b,…,j} classified under two facets. Right side: Reduced taxonomy.

4 Study case

The data set that will be studied consists of a total of 500 artworks [7], [8] (An example can be appreciated on Fig. 3) with description in the Spanish language, of which a training set of 200 artworks was selected for testing.

In order to implement the dynamic taxonomy we used the following facets: genre, connotation and denotation. Genre refers to each one of the different categories or classes under which the artworks can be classified according to common forms and contents. For example: portraits, self-portraits, landscapes, religion, mythology etc. Of these three facets, only genre is defined by an expert user, while the other two are obtained automatically.



Autorretrato con modelo

En este estudio de la **vanidad** decadente, el artista pinta su autorretrato, mientras una **mujer desnuda** yace provocadora detrás de él, físicamente cercana pero emocionalmente distante. El pintor viste una prenda transparente que parece una capa de piel, unida mediante puntadas en el pecho. La mejilla de la **mujer** está surcada por una cicatriz impropia. Detrás de ella se alza una **flor** fálica, un simbólico **narciso**, y más allá de ambos podemos entrever los tejados de la sombría **ciudad** industrial.

Fig. 3. An artistic object with their textual description (denotation).

5 Methodology

In the following paragraphs we will show the steps taken in order to generate a classification system based on dynamic taxonomies.

5.1 Obtaining a Controlled Vocabulary

A Controlled Vocabulary (CV) is an organized lists of words and phrases that are used to initially tag content, and then to find it through navigation or search.

5.1.1 Elimination of stopwords

The first step consists in the elimination of stopwords, since such words do not allow for discrimination of relevant attributes of the objects. There are several lists of stopwords in Spanish (Snowball[1]; Ranks[2]).

5.1.2 Stemmer

In order to obtain a dynamic taxonomy, one must have a controlled vocabulary. In this particular case we applied a Spanish stemmer[3] based on Porter's algorithm, which allows for the decrement of the controlled vocabulary and augments the definition of each concept, providing a better recall.

5.1.3 N-Grams

With the objective of finding concepts formed by more than a word, we obtained bigrams, trigrams and 4-grams of the textual description of artworks. And also, we have a set of words (unigrams).

[1] Snowball, *Spanish stop word list*, http://snowball.tartarus.org/algorithms/spanish/stop.txt
[2] Ranks NL, *Spanish stopwords*, http://www.ranks.nl/stopwords/spanish.html
[3] Snowball, *Spanish stemming algorithm*, http://snowball.tartarus.org/algorithms/spanish/stemmer.html.

5.2 Defining the Concepts

A concept is a label which identifies a set of documents[4] (classified under that concept), every concept is related with a certain level of abstraction that depends with the level in the taxonomy, this make clear that concepts are not terms. So, the problem here is how we can know what terms of phrases of our controlled vocabulary can be a concept, to answer that question we use a thesaurus, using a thesaurus we can define the part of our controlled vocabulary that we can use as concepts as we can see in the Fig. 4 and in the equation 1.

$$Concepts = Thesaurus \cap CV \quad . \tag{1}$$

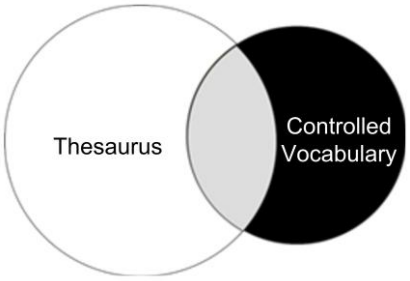


Fig. 4. A Concept definition

5.2.1 Incrementing and Expanding Concepts

By using clustering algorithms, we look for terms that were not originally considered part of the concepts using the remaining concepts of the thesaurus. Afterwards, new concepts are added using a supervised process like the one shown in Fig. 5. Also, there is a possibility to expand the definition of each concept, this leads us to the generation of new clusters between the controlled vocabulary and the concepts.

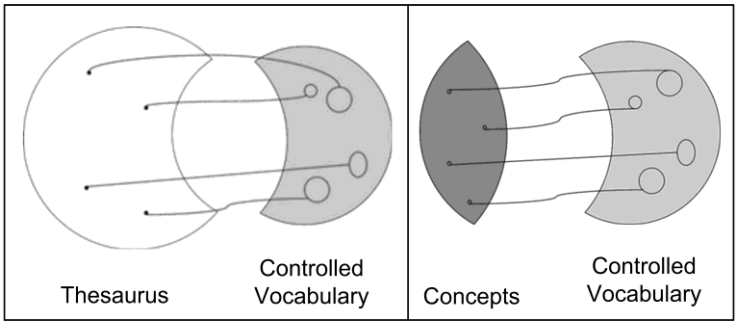


Fig. 5. Left side: Expanding the concepts. Right side: Concept expansion.

[4] Information Management Unit, *Guided Interactive Discovery of e-Government Services*, http://www.imu.iccs.gr/sweg/presentations/Giovanni%20Maria%20Sacco.ppt

5.3 Obtaining taxonomies

The next step consists of generating the taxonomies of the concepts using the thesaurus hierarchies.

5.3.1 Defining the Taxonomy

In this step we obtain the hierarchical structure of every concept based on the thesaurus structure, we use a hash table to do this process (See Fig. 6). The taxonomy is then created by the process of linking all the concepts.

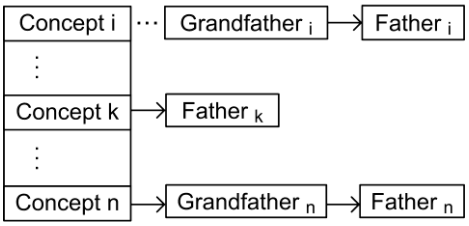


Fig. 6. Obtaining the hierarchy of every concept.

5.3.2 Structuring the Facets

Once the hierarchic structure of the concepts contained in the vocabulary is ready, one must supervise under which facet they will be placed. In Fig 7 we show a simple example of the taxonomy for the facets.

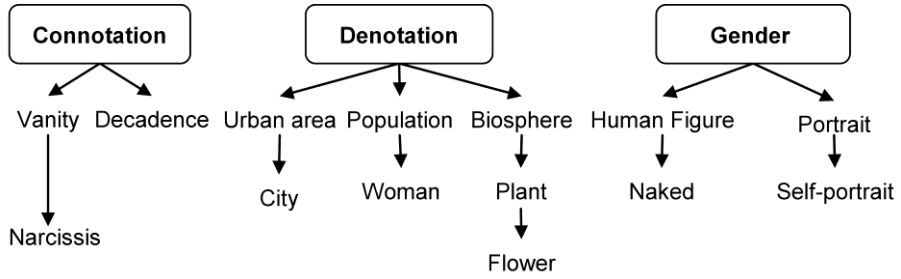


Fig. 7. A facet taxonomy example.

5.3.3 Pruning

The taxonomy should be pruning, to avoid that the user spend time expanding unnecessary nodes, this nodes are those that doesn't helps us in the filter process.

5.3.4 Frequency filter

If we order the controlled vocabulary based on its frequency, it is possible to eliminate the less frequent concepts, due to the fact that they will generally not be used for information recovery. However, our proposal consists in implementing this filter directly over the taxonomy. This process entails a second pruning over the faceted taxonomies.

5.4 Indexing

Each artwork contains a textual description. Such description has words or phrases that are included in the controlled vocabulary. At first, and before we have used hierarchies to bond the controlled vocabulary to the faceted taxonomy, the indexes are not related among each other (Fig. 8).

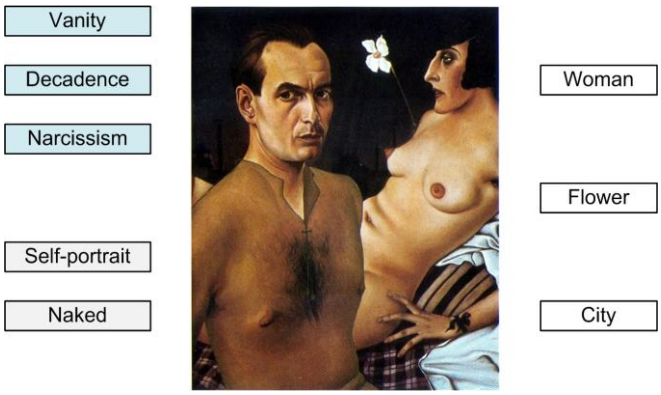


Fig. 8. Denotation index.

However, once created the taxonomies, each index is related to each other by a hierarchical scheme, as shown in Fig. 9. This bonding allows to index the artwork under the information that contains its textual description, and to add the hierarchical information of each concept (Fig. 10).

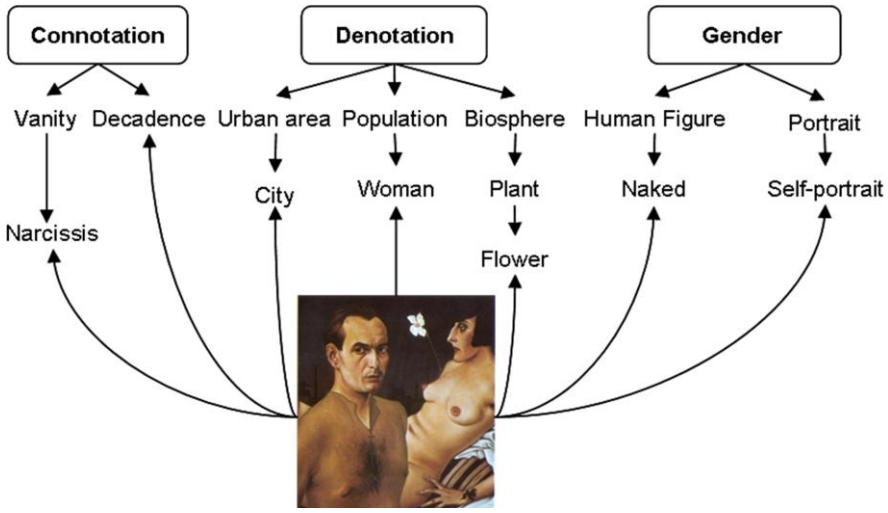


Fig. 9. Concepts connections.

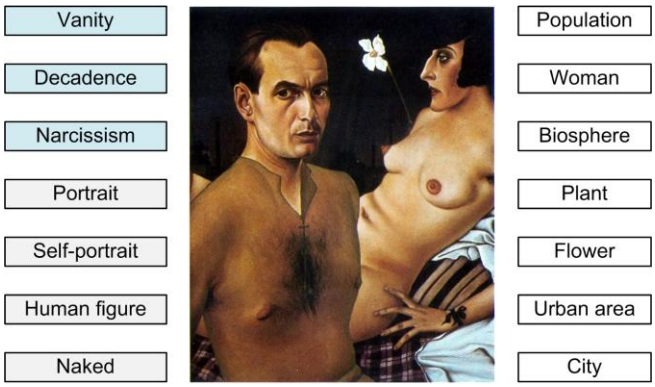


Fig. 10. Conceptual index.

5.5 Storing model

In order to use the model, the facet taxonomy and the objects should be stored, we use the Extended Entity-Relationship diagram shown in Figure 11 for this propose.

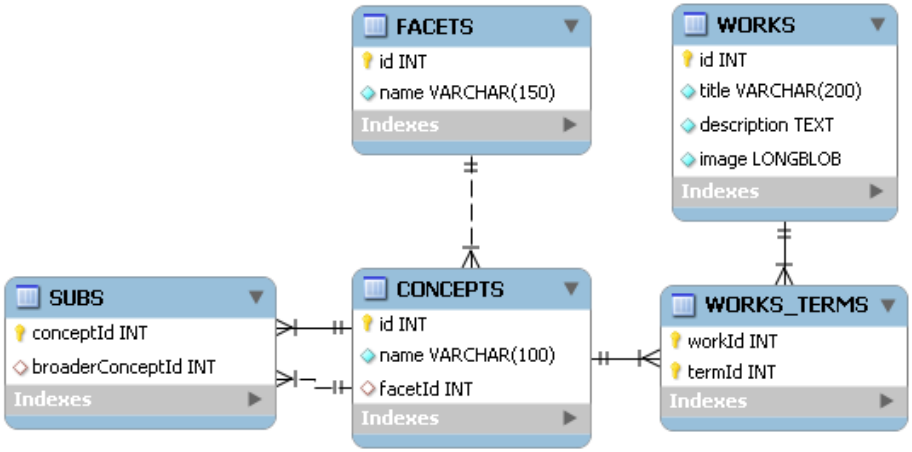


Fig. 11. EER Model for store a facet taxonomy and artworks objects**.**

5.6 Implementation (Navigation Tree)

The final step is make a visual framework in which the user can select and combine appropriate concepts [2], this is develop by a Navigation Tree [9] (taxonomic tree [1]). The navigation tree contains nodes that enable the user to start browsing in one facet and then cross to another, and so on, until reaching the desired level of specificity [9].

6. Results and Evaluation

By using the procedure described earlier we obtained a controlled vocabulary of 500 concepts that describe the 200 documents. It is important to mention that the eliminated stopwords represented 50% of the total words. By means of this procedure we were able to bond each artwork to an average of fourteen indexes.

We performed a comparison between our faceted classification system and the "Full Text Search" function of MySQL, which uses a Boolean search for any given data set. We configured a set of supervised consults, taking into account two criteria: recall and precision. Fig. 12 shows the results.

As one can see in Fig. 12, the faceted classification system provides a significantly higher degree of recall when compared to the Boolean search, underlining the advantage of using a conceptual search instead of a textual search.

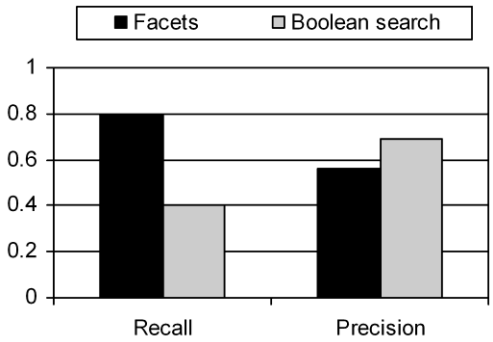


Fig. 12 Recall and Precision evaluation.

7. Conclusions

So far, the automatic solutions for hierarchy constructions available have not offered satisfactory outcomes when faced with the construction of taxonomies. However, some methodologies developed for the construction of facets and the generation of taxonomies based on textual analysis has yielded encouraging results. As of now we are implementing new algorithms that will allow the generation of interpretations based on generic descriptions, expanding the "connotation" facet described in this paper. We have also considered the usage of a terminological conceptual thesaurus for this task, and the possibility of allowing the user to use "open taxonomies" thus allowing the taxonomy to evolve, reflecting the progression of the words and their interpretation and keeping its novelty.

References

1. Patrick Lambe, Organising Knowledge: Taxonomies, Knowledge and Organisational Effectiveness, ISBN: 9781843342274, Chandos Publishing (Oxford) Limited. UK, 2007.
2. G.M. Sacco, Dynamic Taxonomies: A Model for Large Information Bases. IEEE Transactions on Knowledge and Data Engineering 12, 2, pp. 468-479, May 2000.
3. G.M. Sacco, Some Research Results in Dynamic Taxonomy and Faceted Search Systems. SIGIR'2006 Workshop on Faceted Search, August 2006 Seattle, WA, USA.
4. Spangler, S. Kreulen, J.T. Lessler, J. "MindMap: utilizing multiple taxonomies and visualization to understand a document collection", . Proceedings of the 35th Annual Hawaii International Conference on System Sciences, 2002. HICSS. Pags. 1170-1179.
5. E. Stoica and M. Hearst, "Demonstration: Using WordNet to Build Hierarchical Facet Categories". The ACM SIGIR Workshop on Faceted Search, August, 2006
6. Categories for the Description of Works of Art (CDWA), editado por Murtha Baca and Patricia Harpring, http://www.getty.edu/research/conducting_research/standards/cdwa/index.htmll. 2009.
7. El ABC del Arte del siglo XX, Primera edición en español 1999, Editorial Phaidon Press Limited.
8. Masdearte.com, Portal de arte contemporáneo, http://www.masdearte.com/item_critica.cfm?id=315. 2009.
9. Y. Tzitzikas, A. Analyti, N. Spyratos and P. Constantopoulos, An Algebra for Specifying Valid Compound Terms in Faceted Taxonomies, Journal on Data and Knowledge Engineering (DKE), 62(1), 2007.

The Use of Document Fingerprinting in the Web People Search Task*

David Pinto, Mireya Tovar, Beatriz Beltrán, Darnes Vilariño, Héctor Furlog

Faculty of Computer Science, BUAP
14 Sur & Av. San Claudio, CU, Edif. 104C
Puebla, Mexico, 72570
{dpinto, mtovar, bbeltran, darnes}@cs.buap.mx
http://nlp.cs.buap.mx

Abstract. In the context of document indexing/retrieval, a document fingerprint is considered to be a specific code which may be used to uniquely identify this document from the rest of the text collection. Document fingerprinting is a efficient time-complexity mechanism of indexing data, but issues with respect to precision still being on development. In this paper, we approached the Web People Search task (WePS) by using hash-based document fingerprinting. The evaluation of the experiments carried out show that the implemented technique could have a positive impact in the analysis/indexing of huge volumes of information. However, the feature set for all the documents in the WePS framework needs to be further investigated.

1 Introduction

Classifying people names on the web is a task that requires a special attention by the classification societies community. Searching people in Internet is one of the most common activities performed by the World Wide Web users [1]. The main challenge consists in bringing together all the results (of a given search engine) that share the same occupation/profession (which very often are ambiguous) by using a highly scalable classification method.

In this paper, we report the results obtained in the Web People Search task [2] when using a system based on hash-based text retrieval techniques [3]. In particular, we have constructed a new vectorial coordinates system for the representation of the original data and, thereafter, we calculate the distance of the vectorial representation of each input dataset by means of a hash function.

The experiments were carried out by using the WePS-2 collection. In summary, it is made up of 30 ambiguous names of people, each name with a number of html pages with information related with that people name. The complete description of the evaluated corpus is given into detail in [2].

We must take into account that the fingerprinting technique may allow indexing and clasifying of documents in a one single step. Therefore, given the huge

* This work has been partially supported by the CONACYT project #106625, as well as by the PROMEP/103.5/09/4213 grant.

amount of information available in Internet, we consider that the main contribution of this research work consists of providing a very fast way of classifying people names on the World Wide Web.

The evaluation of the experiments carried out show that the implemented technique could have a positive impact in the analysis/indexing of huge volumes of information. However, the feature set for all the documents in the WePS framework needs to be further investigated.

The remainder of this document is structured as follows. Section 2 presents the document fingerprint technique used in the process of indexing and clustering of documents in the Web People Search framework. In Section 3 we describe the components of the implemented system. The experimental results are discussed in Section 4. Finally in Section 5 the conclusions are given.

2 Document Fingerprinting

Document indexing based on fingerprinting is a powerful technology for similarity search in huge volumes of documents. The goal is to provide a proper hash function which cuasi-uniquely identifies each document, so that the hash collisions may be interpreted as similarity indication.

Formally, given two documents d_1 and d_2, and the fingerprint of the two documents $h(d_1)$ and $h(d_2)$, respectively. We consider d_1 and d_2 to be ϵ-similar iff $|h(d_1) - h(d_2)| < \epsilon$.

In the context of document indexing/clustring/retrieval a fingerprint $h(d)$ of a document d may be considered as a set of encoded substrings taken from d, which serve to identify d uniquely.

Defining the specific hash function to encode the substrings of the documents is the main challenge of the fingerprinting technique. In particular, in the implementation of the BUAP Web People Search system we defined a small set of term-frequency vectors (which are used as reference for a new system coordinates) in order to be considered as the new reference for the vectorial representation of each document of the WePS-2 collection. In Figure 1 we may see an overview of the proposed approach.

Formally, given a set of k reference vectors, $\{\boldsymbol{r_1}, \boldsymbol{r_2}, \cdots, \boldsymbol{r_k}\}$, and the vectorial representation of a document d. We defined the fingerprint of $\boldsymbol{d}$ as shown in equation (1).

$$h(\boldsymbol{d}) = \sum_{i=1}^{k} \boldsymbol{r_1} \cdot \boldsymbol{d} \tag{1}$$

The specific features used in the vectorial representation of the documents are explained in the following section.

3 The BUAP system

The system comprises the following components:

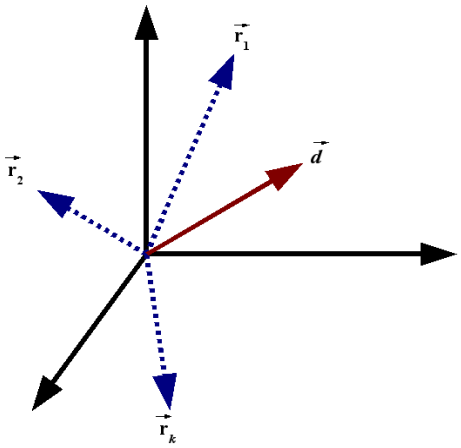


Fig. 1. The new coordinates system used in the implemented hash-based function for fingerprinting.

Pre-processing: We have programmed two implementations in order to perform the HTML to text conversion. The first HTML to text converter was programmed with Java, whereas the second was implemented with AWK. No HTML tags nor url's were considered in the text extraction.

Named entity recognition: We used the Stanford Named Entity Recognizer [4] in order to extract names of places, organizations and people names from the target documents.

Document representation: The features used to represent each document comprised all the named entities recognized by the Stanford NER. Thus, the vectorial representation of a document d was:

$$\boldsymbol{d} = \{tf(ne_1), tf(ne_2), \cdots, tf(ne_n)\}, \tag{2}$$

where $tf(ne_i)$ is the frequency of the i-th named entity recognized in the document d.

Reference vector generation: In order to generate the k reference vectors, we calculated the named entity vocabulary, V_{NE}, of the entire target collection. We sorted this vocabulary in a non-increased order according to the named entity frequency (over the complete collection) and, thereafter, we selected only those named entities whose frequency were between the so called "transition" range [5]. The transition range allows to obtain the mid-frequency terms of a given vocabulary.

A typical formula used to obtain the center of the transition range (*transition point*) is given in Equation (3).

$$TP_V = \frac{\sqrt{8 * I_1 + 1} - 1}{2} \tag{3}$$

where I_1 represents the number of terms (in our case, named entities) with frequency equal to 1.
Once the *transition point* has been found, we may extract the mid-frequency named entities, which are those which obtain the closest frequency values to TP_V, i.e.,

$$V_{TP} = \{ne_i | ne_i \in V_{NE}, U_1 \leq tf_c(ne_i) \leq U_2\}, \quad (4)$$

where $tf_c(ne_i)$ is the frequency of the i-th entity over the complete document collection and U_1 is a lower threshold obtained by a given neighbourhood value of the TP: $U_1 = (1 - NTP) * TP_V$, where $0 \leq NTP < 1$. U_2 is the upper threshold and it is calculated in a similar way: $U_2 = (1 + NTP) * TP_V$. Thus, the representation of the j-th reference vector $\boldsymbol{r_j}$ is given as follows:

$$\boldsymbol{r_j} = \{tf_c(ne_1), tf_c(ne_2), \cdots, tf_c(ne_m)\}. \quad (5)$$

Indexing/clustering: The indexing process was carried out by using the formula expressed in Equation (1). We used a specific threshold (ϵ) in order to determine a range of hash-based values (documents) that should belong to the same cluster. The overlapping of clusters was not considered but it may be easily implemented.

4 Experimental results

Besides the evaluation of the WePS-2 collection, we performed a set of experiments over the training and test dataset of the WePS-1 collection (see [1] for a complete description of these datasets). The obtained results are presented in Tables 1, 2 and 3, respectively.

In these tables we may see the following set of metrics used to evaluate the performance of the implemented system:

BEP: BCubed Precision
BER: BCubed Recall
FMeasure_0.5_BEP-BER: F-measure of B-Cubed P/R with alpha set to 0.5
FMeasure_0.2_BEP-BER: F-measure of B-Cubed P/R with alpha set to 0.2
P: Purity
IP: Inverse Purity
FMeasure_0.5_P-IP: F-measure of Purity and Inverse Purity with alpha set to 0.5
FMeasure_0.2_P-IP: F-measure of Purity and Inverse Purity with alpha set to 0.2

For more details about the evaluation metrics please refer to [6]. The baselines and the rationale for F-measures with alpha 0.2 are explained in the WePS-1 task description paper [1].

We have tested six differents approaches varying the document similarity threshold (ϵ), the HTML to text converter and the use or not of named entities. The name of each approach as well as their describtion is given as follows:

BUAP_1 ($\epsilon = 0.0004$): Java-based HTML to text converter without NER, i.e., all the document terms are used.
BUAP_2 ($\epsilon = 0.0004$): Java-based HTML to text converter with NER, i.e., all the document named entities are used.
BUAP_3 ($\epsilon = 0.0004$): AWK-based HTML to text converter without NER, i.e., all the document terms are used.
BUAP_4 ($\epsilon = 0.0004$): AWK-based HTML to text converter with NER, i.e., all the document named entities are used.
BUAP_5 ($\epsilon = 0.3$): Java-based HTML to text converter without NER, i.e., all the document terms are used.
BUAP_6 ($\epsilon = 0.3$): AWK-based HTML to text converter without NER, i.e., all the document terms are used.

We may see (Tables 1, 2 and 3) that the implemented approaches obtained a performance comparable with two of the proposed baselines, ALL_IN_ONE and ONE_IN_ONE, with a document similarity threshold (ϵ) equal to 0.3 and 0.0004, respectively. The selection of any of the two HTML to text converters was not important with respect to the obtained results. Moreover, we could not confirm the benefit of using named entities with respect to those approaches that did not use them, because the obtained results did not show significant difference among the different approaches.

Although, some of the implemented approaches obtained acceptable results in comparison with the baselines, in the case of the WePS-2 collection any of the first four approaches outperformed the proposed baselines. We consider that the expected document distribution over the final clusters has played an important role on the obtained results, since the presented algorithm of fingerprinting usually assumes a uniform distribution of documents over the discovered clusters.

The evaluation of the experiments carried out show that the implemented technique could have a positive impact in the analysis/indexing of huge volumes of information. However, the feature set for all the documents in the WePS framework needs to be further investigated.

As future work, we would like to experiment on feature selection in order to clearly benefit the construction of the reference vector set. Although we would like to keep the process as unsupervised as possible, we are considering the use of supervised classifiers in order to extract the most important features to tackle the Web People Search task.

Finally, we would like to analyse the use new hash-based functions and new document representations which consider characteristics other than only term or named entity frequencies.

5 Conclusions

We implemented a hash-based function in order to uniquely identify each document from a text collection in the framework of the Web People Search task. The hash collisions were interpreted as similarity degree among the target documents. In this way, we constructed an algorithm which only takes into account

Table 1. Evaluation of the WePS-2 test dataset.

run	BEP	BER	FMeasure_0.2 BEP-BER	FMeasure_0.2 P-IP	FMeasure_0.5 BEP-BER	FMeasure_0.5 P-IP	IP	P
ALL_IN_ONE_BASELINE	0.43	1.0	0.66	0.79	0.53	0.67	1.0	0.56
COMBINED_BASELINE	0.43	1.0	0.65	0.94	0.52	0.87	1.0	0.78
ONE_IN_ONE_BASELINE	1.0	0.24	0.27	0.27	0.34	0.34	0.24	1.0
BUAP_1	0.89	0.25	0.27	0.30	0.33	0.37	0.27	0.89
BUAP_2	0.89	0.24	0.27	0.29	0.33	0.35	0.26	0.89
BUAP_3	0.89	0.24	0.27	0.29	0.33	0.36	0.26	0.89
BUAP_4	0.90	0.24	0.27	0.29	0.33	0.36	0.26	0.90
BUAP_5	0.44	1.00	0.67	0.80	0.53	0.67	1.00	0.56
BUAP_6	0.44	1.00	0.66	0.80	0.53	0.67	1.00	0.56

Table 2. Experimental results with the training dataset of the WePS-1 collection.

run	BEP	BER	FMeasure_0.5 BEP-BER	FMeasure_0.5 P-IP	IP	P
ALL_IN_ONE_BASELINE	0.54	1.0	0.64	0.75	1.0	0.65
ONE_IN_ONE_BASELINE	1.0	0.34	0.45	0.46	0.35	1.0
COMBINED_BASELINE	0.48	1.0	0.60	0.9	1.0	0.82
BUAP_1	0.64	0.6	0.56	0.55	0.68	0.54
BUAP_2	0.64	0.6	0.56	0.55	0.68	0.54
BUAP_3	0.6	0.69	0.58	0.56	0.76	0.51
BUAP_4	0.6	0.69	0.58	0.56	0.76	0.51
BUAP_5	0.57	0.93	0.65	0.61	0.96	0.50
BUAP_6	0.54	0.99	0.65	0.61	1.00	0.48

Table 3. Experimental results with the test dataset of the WePS-1 collection.

run	BEP	BER	FMeasure_0.5 BEP-BER	FMeasure_0.5 P-IP	IP	P
ALL_IN_ONE_BASELINE	0.18	0.98	0.25	0.4	1.0	0.29
COMBINED_BASELINE	0.17	0.99	0.24	0.78	1.0	0.64
ONE_IN_ONE_BASELINE	1.0	0.43	0.57	0.61	0.47	1.0
BUAP_1	0.27	0.61	0.31	0.36	0.71	0.29
BUAP_2	0.27	0.61	0.31	0.36	0.71	0.29
BUAP_3	0.23	0.73	0.28	0.38	0.82	0.29
BUAP_4	0.23	0.73	0.28	0.38	0.82	0.29
BUAP_5	0.21	0.92	0.30	0.36	0.96	0.25
BUAP_6	0.18	0.98	0.25	0.36	1.00	0.25

the local features of each document in order to index/cluster them. The experimental results over the WePS-1 test and training datasets showed an acceptable performance of the proposed algorithm. However, the proposed reference vector for the fingerprinting-based model were useless when evaluating with the WePS-2 dataset. The proper construction of reference vectors for the automatic and unsupervised classification of people names in the Web needs to be further investigated.

References

1. Artiles, J., Gonzalo, J., Sekine, S.: The SemEval-2007 WePS evaluation: Establishing a benchmark for the web people search task. In: Proc. of the 4th International Workshop on Semantic Evaluations - SemEval 2007, Association for Computational Linguistics (2007) 64–69
2. Artiles, J., Gonzalo, J., Sekine, S.: Weps 2 evaluation campaign: overview of the web people search clustering task. In: Proc. of the 2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference. (2009)
3. Stein, B.: Principles of hash-based text retrieval. Clarke, Fuhr, Kando, Kraaij, and de Vries, Eds., 30th Annual Int. ACM SIGIR Conf. (2007) 527–534
4. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by gibbs sampling. In: Proceedings of the ACL. (2005) 363–370
5. Pinto, D., Jiménez-Salazar, H., Rosso, P.: Clustering abstracts of scientific texts using the transition point technique. In: Proc. of the CICLing 2006 Conference. Volume 3878 of Lecture Notes in Computer Science., Springer-Verlag (2006) 536–546
6. Amigó, E., Gonzalo, J., Artiles, J., Verdejo, F.: A comparison of extrinsic clustering evaluation metrics based on formal constraints. Information Retrieval **12**(4) (2009) 461–486

mQA: Question Answering in Mobile devices

Fernando Zacarías F.[1], Alberto Tellez V.[2], Marco Antonio Balderas[3], and Rosalba Cuapa C.[4]

Benemérita Universidad Autónoma de Puebla,
[1,3,4]Computer Science and [2] Collaborator - INAOE
14 Sur y Av. San Claudio, Puebla, Pue.
72000 México
[1]fzflores@yahoo.com.mx, [2]albertotellezv@ccc.inaoep.mx
[3]balderasespmarco@gmail.com, [4]rcuapa_canto@yahoo.com

Abstract. In this paper, we present a novel proposal for Question Answering through mobile devices. Thus, an architecture for a mobile Question Answering system based on WAP technologies is deployed. The architecture propose moves the issue of Question Answering to the context of mobility. This paradigm ensures that QA is seen as an activity that provides entertainment and excitement pleasure. This characteristic gives to QA an added value. Furthermore, the method for answering definition questions is very precise. It could answer almost 90% of the questions; moreover, it never replies wrong or unsupported answers. Considering that the mobile-phone has had a boom in the last years and that a lot of people already have mobile telephones (approximately 3.5 billions), we propose an architecture for a new mobile system that makes QA something natural and effective for work in all fields of development. This obeys to that the new mobile technology can help us to achieve our perspectives of growth. This system provides to user with a permanent communication in anytime, anywhere and any device (PDA's, cell-phone, NDS, etc.).

Keywords: Mobile devices, Question Answering, WAP, GPRS.

1 Introduction

Each generation of mobile communications has been based on a dominant technology, which has significantly improved spectrum capacity. Until the advent of IMT-2000, cellular networks had been developed under a number of proprietary, regional and national standards, creating a fragmented market.

- First Generation was characterized for Advanced Mobile Phone System (AMPS). It is an analog system based on FDMA (Frequency Division Multiple Access) technology. However, there were also a number of other proprietary systems, rarely sold outside the home country.
- Second Generation, it includes five types of cellular systems mainly:

  - Global System for Mobile Communications (GSM) was the first commercially operated digital cellular system.
  - GSM uses TDMA (Time Division Multiple Access) technology.
  - TDMA IS-136 is the digital enhancement of the analog AMPS technology. It was called D-AMPS when it was fist introduced in late 1991 and its main objective was to protect the substantial investment that service providers had bmade in AMPS technology.
  - CDMA IS-95 increases capacity by using the entire radio band with each using a unique code (CDMA or Code Division Multiple Access)
  - Personal Digital Cellular (PDC) is the second largest digital mobile standard although it is exclusively used in Japan where it was introduced in 1994.
  - Personal Handyphone System (PHS) is a digital system used in Japan,
- Third Generation, better known as 3G or 3rd Generation, is a family of standards for wireless communications defined by the International Telecommunication Union, which includes GSM EDGE, UMTS, and CDMA2000 as well as DECT and WiMAX. Services include wide-area wireless voice telephone, video calls, and wireless data, all in a mobile environment. Thus, 3G networks enable network operators to offer users a wider range of more advanced services while achieving greater network capacity through improved spectral efficiency.

Currently, mobile devices are part of our everyday environment and consequently part of our daily landscape [5]. The current mobile trends in several application areas have demonstrated that training and learning no longer needs to be classroom. Current trends suggest that the following three areas are likely to lead the mobile movement: m-application, e-application and u-application. There are estimated to be 2.5 billion mobile phones in the world today. This means that this is more than four times the number of personal computers (PCs), and today's most sophisticated phones have the processing power of a mid-1990s PC. Even, in a special way, many companies, organizations, people and educators are already using iPhone, iPod, NDS, etc., in their tasks and curriculas with great results. They are integrating audio and video content including speeches, interviews, artwork, music, and photos to bring lessons to life. Many current developments, just as ours [5, 3, 6], incorporate multimedia applications.

In the late 1980's, a researcher at Xerox PARC named Mark Weiser [4], coined the term "Ubiquitous Computing". It refers to the process of seamlessly integrating computers into the physical world. Ubiquitous computing includes computer technology found in microprocessors, mobile phones, digital cameras and other devices. All of which add new and exciting dimensions to applications.

As pragmatic uses grow for cellphones, mobile technology is also expanding into creative territory. New public space art projects are using cellphones and other mobile devices to explore new ways of communicating while giving everyday people the chance to share some insights about real world locations.

While your cellphone now allows you to play games, check your e-mail, send text messages, take pictures, and oh, yeah, make phone calls, it can perhaps serve a more enriching purpose. Thus, we think that widespread internet access and collaboration technologies are allowing businesses of all sizes to mobilise their workforce. Such innovations provide additional flexibility without the need to invest in expensive and complex on-premise infrastructure requirements. Furthermore, it makes "eminent sense" to fully utilise the web commuting options provided by mobile technology.

The problem of answering questions has been recognized and partially tacled since the 70's for specific domains. However, with the advent of browsers working with billions of documents in internet, the need has newly emerged, having led to approaches for open-domain QA. Some examples of such approaches are emergent question answering engines such as *answers.com*, *ask.com*, or additional services in traditional nrowsers, such as *Yahoo*.

Recent research in QA has been mainly fostered by the TREC and CLEF conferences. The first one focus on English QA, whereas the second evaluates QA systems for most European languages except English. To do, both evaluation conferences have considered only a very restriced version of the general QA problem. They basically contemplate simple questions which assume a definite answer typified by a named entity or noun phrase, such as factoid questions (for instance, "How old is Cher?" or "Where is the Taj Mahal?") or definition questions ("Who is Nelson Mandela?" or "What is the quinoa?"), and exclude complex questions such as procedural or epaculative ones.

Our paper is structured as follows: In section 2 we describe the state of the art about QA and similar works. Next, in section 4 we present our mobile architecture to support question answering. Section 5 contains our perspectives about our future work. This work consist in incorporate answering definitions questions. Finally, the conclusions are drawn in section 6.

2 The state of the art

One of the oldest problems of human history is raising questions about several issues and conflicts that torments our existence. Since children this is the mechanism we use to understand and adapt to our environment. The counterpart to ask questions is to answer the questions that we do, an activity that also requires intelligence. This activity has a difficulty level that has tried to delegate to computers, almost since the emergence of these. The issue of question answering for a computer has been recognized and tackled from the decade of the 70s century past for specific domains. In Mexico, have been obtained excellent results in this context, for this reason we propose to bring these same results with mobile technologies.

Recent research has focused on developing systems for question answering to open domain, ie systems that takes as their source of information a collection of texts on a variety of topics, and solve questions whose answers can be obtained from the collection of departure. From question answering systems developed so far, we can identify three main phases:

1. *Analysis of the question.* This first phase will identify the type of response expected from the given question, that is expected to be a question of "when" a kind of response time, or a question "where" will lead us to identify a place. Response rates are most commonly used personal name, name organization, number, date and place.
2. *Recovery of the document.* In the second stage performs a recovery process on the collection of documents using the question, which is to identify documents on the question that probably contain the kind of response expected. The result of this second stage is a reduced set of documents and preferably specific paragraphs.
3. *Extraction of the response.* The last phase uses the set of documents obtained in the previous phase and the expected type of response identified in the first phase, to locate the desired response.

Questions of definition require a more complex process in the third stage, since they must obtain additional information segments and at the same time are not repetitive. To achieve a good "definition" must often resort to various documents [1].

Currently the question answering on mobile devices for open domains is in a development stage. The project QALL-ME, is a project of 36 months, funded by the European Union and will be conducted by a consortium of seven institutions, including four academic and three industrial companies. The aim is to establish a shared infrastructure for developing a QA infrastructure via mobile phone for any tourist or citizen can instantly access to different information regarding the services sector, be it a movie in the cinema, a theater or restaurant of a certain type of food. All this in a multilingual and multimodal mode for mobile devices. The project will experiment with the potential of open domain QA and evaluation in the context of seeking information from mobile devices, a multimodal scenery which includes natural speech as input, and the integration of textual answers, maps, pictures and short videos as output.

The architecture proposed in the QALL-ME project is a distributed architecture in which all modules are implemented as Web services using standard language for defining services. In figure 1 shows the main modules of this architecture. The architecture of the QALL-ME described as follows:

"The central planner is responsible for interpreting multilingual queries. This module receives the query as input, processes the question in the language in which it develops and, according to the parameters of context, directs the search for required information. Extractor to a local response. The extraction of the

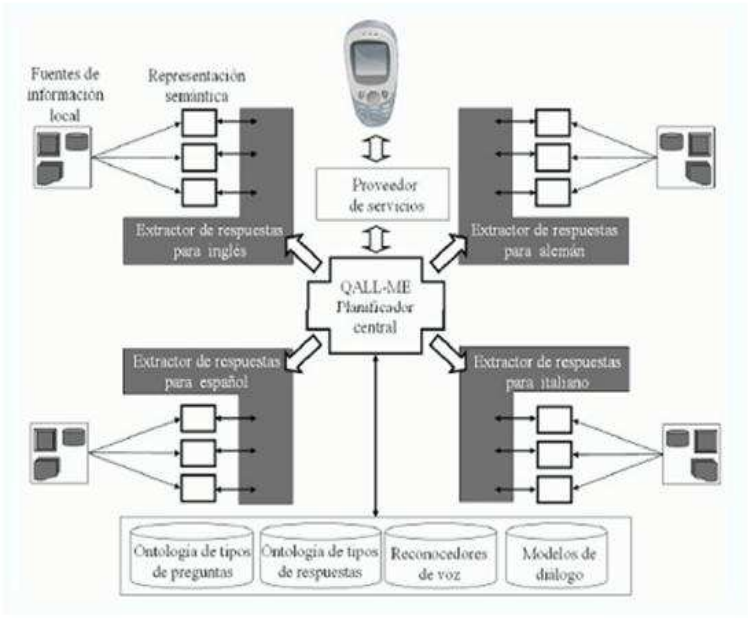


Fig. 1. Main QALL-ME Architecture [8]

response is made on different semantic representations of the information depends on the type of the original source data from which we get the answer (if the source is plain text, the semantic representation is an annotated XML document if the source is a website, the semantic representation is a database built by a wrapper). Finally, the responses are returned to the central planners to determine the best way to represent the requested information" [8].

3 Movile Question Answering for Definitions Questions

The method for answering definition questions uses Wikipedia [10] as target document collection. It takes advantage of two known facts: [10] Wikipedia organizes information by topics, that is, each document concerns one single subject and, [11] the first paragraph of each document tend to contain a short description of the topic at hand. This way, it simply retrieves the document(s) describing the target term of the question and then returns some part of its initial paragraph as answer.

Figure 2 shows the general process for answering definition questions. It consists of three main modules: target term extraction, document retrieval and answer extraction.

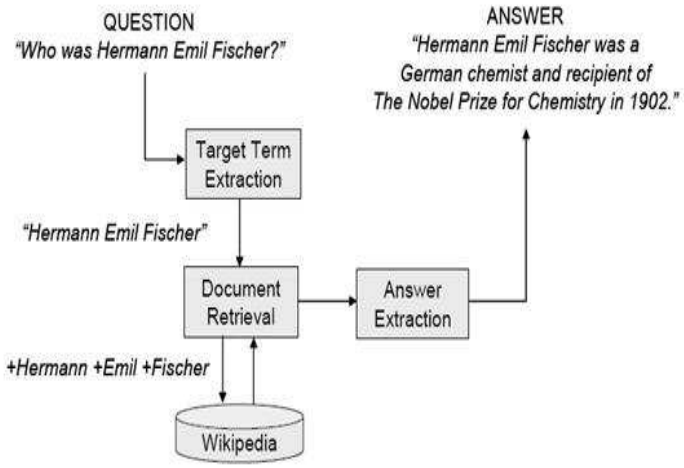


Fig. 2. Process for answer definition questions [7]

3.1 Finding Relevant Documents

In order to search in Wikipedia for the most relevant document to the given question, it is necessary to firstly recognize the target term. For this purpose the method uses a set of manually constructed regular expressions such as: "What—Which—Who—How"+"any form of verb to be"+<TARGET>+"?", "What is a <TARGET> used for?", "What is the purpose of <TARGET>?", "What does <TARGET> do?", etc. Then, the extracted target term is compared against all document names and the document having the greatest similarity is recovered and delivered to the answer extraction module. It is important to mention that, in order to favor the retrieval recall, we decided using the document names instead of the document titles since they also indicate their subject but normally they are more general (i.e., titles tend to be a subset of document names). In particular, the system uses the Lucene [11] information retrieval system for both indexing and searching.

3.2 Extracting the Target Definition

As we previously mentioned, most Wikipedia's documents tend to contain a brief description of its topic in the first paragraph. Based on this fact, this method for answer extraction is defined as follows:

– Consider the first sentence of the retrieved document as the target definition (the answer).

– Eliminate all text between parenthesis (the goal is to eliminate comments and less important information).
– If the constructed answer is shorter than a given specified threshold2, then aggregate as many sentences of the first paragraph as necessary to obtain an answer of the desire size.

For instance, the answer for the question "Who was Hermann Emil Fischer?" (refer to Figure 2) was extracted from the first paragraph of the document "Hermann_Emil_Fischer": "Hermann Emil Fischer (October 9, 1852 - July 15, 1919) was a German chemist and recipient of the Nobel Prize for Chemistry in 1902. Emil Fischer was born in Euskirchen, near Cologne, the son of a businessman. After graduating he wished to study natural sciences, but his father compelled him to work in the family business until determining that his son was unsuitable".

3.3 Evaluation Results of our method

This section presents the experimental results about the participation [7] at the monolingual Spanish QA track at CLEF 2007. This evaluation exercise considers two basic types of questions, definition and factoid. However, this year there were also included some groups of related questions. From the given set of 200 test question, our QA system treated 34 as definition questions and 166 as factoid. Table 3.3 details our general accuracy results.

Table 1. System's general evaluation

	Right	Wrong	Inexact	Unsupported	Accuracy
Definition	30	-	4	-	88.23%
Factoid	39	118	3	6	23.49%
TOTAL	69	118	7	6	34.50%

It is very interesting to notice that our method for answering definition questions is very precise. It could answer almost 90% of the questions; moreover, it never replies wrong or unsupported answers. This result evidenced that Wikipedia has some inherent structure, and that our method could effectively take advantage of it. [7]

4 Proposed Architecture

New technologies such as Wireless Application Protocol (WAP), General Packet Radio Service (GPRS) and 3G (3rd generation) further increase communication technologies and access to information. Wireless Application Protocol (commonly referred as WAP) is an open international standard for application layer network communications in a wireless communication environment. Its main use is to enable access to the Mobile Web from a mobile phone or PDA. The recent advances in mobile technology and wireless basic requirements of the applications of WAP has helped improve trade and mobile services. This protocol is a secure specification allowing users to access services and information instantly through wireless mobile devices. WAP is composed of the following form: uses Wireless Markup Language (WML), which includes the Handheld Device Markup Language (HDML). WML can also trace its roots to eXtensible Markup Language (XML). The best known markup language is Hypertext Markup Language (HTML). Unlike HTML, WML is considered a Meta language. WAP also allows the use of standard Internet protocols such as UDP, IP and XML. Although WAP supports HTML and XML, the WML language (an XML application) is specifically devised for small screens and one-hand navigation without a keyboard.

WAP also supports WML Script. It is similar to JavaScript, but makes minimal demands on memory and CPU power because it does not contain many of the unnecessary functions found in other scripting languages. The WAP programming model is similar to the Web programming model with matching extensions, but it accommodates the characteristics of the wireless environment. Figure 3 illustrates this model.

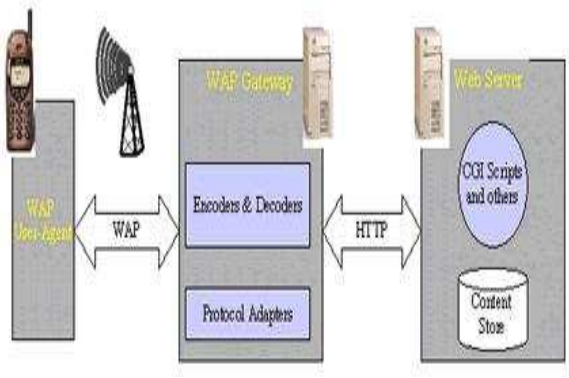


Fig. 3. Rga WAP Programming Model [9]

As you can see, the WAP programming model is based heavily on the Web programming model. But how does the WAP gateway work with HTML? In some cases, the data services or content located on the Web server is HTML-based. Some WAP gateways could be made to convert HTML pages into a format that can be displayed on wireless devices. But because HTML wasn't really designed for small screens, the WAP protocol defines its own markup language, the Wireless Markup Language (WML), which adheres to the XML standard and is designed to enable powerful applications within the constraints of handheld devices. In most cases, the actual application or other content located on the Web server will be native WAP created with WML or generated dynamically using Java servlets or JSP.

In HTML, there are no functions to check the validity of user input or to generate messages and dialog boxes locally. To overcome this limitation, JavaScript was developed. Similarly, to overcome the same restrictions in WML, a new scripting language known as WMLScript has been developed. I'll cover more on WML and WMLScript in later sections. However, General packet radio service (GPRS) is a packet oriented mobile data service available to users of the 2G cellular communication systems, global system for mobile communications (GSM) and provides data rates of 56-114 kbit/s.

GPRS upgrades GSM data services providing:

- Multimedia messaging service (MMS)
- Short message service (SMS)
- Push to talk over cellular (PoC/PTT)
- Instant messaging and presence-wireless village
- Internet applications for smart devices through wireless application protocol (WAP)
- Point-to-point (P2P) service: inter-networking with the Internet (IP)

Some mobile phone operators offer flat rate access to the Internet, while others charge based on data transferred, usually rounded up to 100 kilobytes. During the heyday of GPRS in the developed countries, around 2005, typical prices varied from EUR €0.24 per megabyte to over €20 per megabyte. In developing countries, prices vary widely, and change. Some operators gave free access while they decided pricing, for example in Togocel.tg in Togo, West Africa, others were over-priced, such as Tigo of Ghana at one US dollar per megabyte or Indonesia at $3 per megabyte. AirTel of India charges $0.025 per megabyte, and Telstra of Australia charges $22.53 per megabyte. As of 2008, data access in Canada is still prohibitively expensive. For example, Fido charges $0.05 per kilobyte, or roughly $50 per megabyte. In Venezuela, Digitel charges about $20 per 100 Mb or $25 for unlimited access. In Mexico charges $.04 per Kb. or roughly $40 per megabyte.

This WAP and GPRS infrastructure gives us the appropriate stage to propose the following architecture for QA on mobile devices.

Fig. 4. Main mQA Architecture

In Figure 4 shows the proposed architecture for the definition question answering (DQA), the module will be implemented question answering as Web services using standard languages. It takes a natural language query from a mobile device, this question is sent to the web service which returns a specific answer from a collection of information sources.

The proposed architecture consists of the following modules:

- Mobile Question Answering (mQA): is the interface of the mobile device to the user, which is responsible for communicating via GPRS to the web service DQA.
- Web Service Definition Question Answering (DQA): is the web service to meet all the demands of mQA and from the site, to send all questions to LabTL QA engine to be answered
- "LabTL QA engine: is responsible for seeking the answer to the definition question."
- " Repository of information Wikipedia Spanish-English: Spanish repository in which if the translation to English exists, this will be indicated to the user.

In a broader way, the user on your mobile device generates a question of definition in mQA, the system provides mobile communication system to DQA by GPRS connection (for economy and accessibility by users in case of Mxico 4 cents per KB) , in the DQA service the questions are sent to the LabTL QA system and this searches for the answer in its collection of information (Wikipedia), the answer is validated and this is returned to the user by previously established connection to the mobile device, so you use it according to your needs This architecture pretends to demonstrate the solutions efficiency in question answering by their integration into specific scenarios by mobile devices.

5 Perspectives and Future work

People throughout the world are increasingly relying on cell phones and mobile devices to keep them plugged in. Obviously, search will play an ever increasing role in the evolution of mobile. When will mobile search surpass desktop search? We have been expecting better search capabilities from mobile devices for some time, and know that Asia is far ahead of North America in this respect at the current time. Today, experts discuss their views about the evolution of search in North America. And, what we are sure, is that we must continue working on this line. For this purpose, the next phase of development is the implementation of the Mobile Question Answering System for spanish and English. Furthermore, we seek the application of such search in some opportunity niches such as education.

To sum up the results expected from our architecture presented in this article are:

- Architecture presented here, unlike other proposals based on short text messages [2] is cheaper, such as was presented in section 4.
- Our proposal gives a better performance because the communication via WAP is much more reliable than that based on SMS. This is mainly due to SMS-based systems have a 80 percent certainty. While the WAP protocol provides a 100 percent reliability.
- Our proposal makes use of only a servlet on the server side and a simple midlet on the side of mobile device.
- Furthermore, our proposal will benefit from the availability of Spanish WIKIPEDIA.
- Finally, our proposal is based on Java Micro Edition, thus it will be independent of Operating Systems (OS).

6 Conclusions

A consortium of companies are pushing for products and services to be based on open, global standards, protocols and interfaces and are not locked to proprietary technologies. The architecture framework and service enablers will be independent of Operating Systems (OS). There will be support for interoperability of applications and platforms, seamless geographic and intergenerational roaming. Mobile archutecture proposed in this paper has the advantage of being adaptable to any system and infrastructure, following the current trend that mobile technologies demand.

We believe the selection of topics covered in encyclopedias like WIKIPEDIA for a language is not universal, but reflects the salience attributed to themes in a particular culture that speaks the language. Our approach also would benefit from the availability of the Spanish WIKIPEDIA and the English WIKIPEDIA.

Acknowledgments
Thank you very much to the Autonomous University of Puebla for their financial support. This work was supported under project VIEP register number 15968. Also, we thank the support of the academic body: Sistemas de Informacin.

References

1. A. Lopez. La busqueda de respuestas, un desafio computacional antiguo y vigente. La jornada de Oriente http://ccc.inaoep.mx/cm50-ci10/columna/080721.pdf, 1(1):1-2, July 2008.
2. L. Jochen, The Deployment of a mobile question answering system. Search Engine Meeting. Boston, Massachusetts, 1(1), April 2005.
3. F. Zacaras Flores, F. Lozano Torralba, R. Cuapa Canto, A. Vzquez Flores. English's Teaching Based On New Technologies. The International Journal of Technology, Knowledge & Society, Northeastern University in Boston, Massachussetts, USA. ISSN: 1832-3669, Common Ground Publishing, USA 2008.
4. Weiser, M. (1991). The computer for the twenty-first century. Scientific American, September, 94-104.
5. Zacarías F., Sánchez A., Zacarías D., Méndez A., Cuapa R. FINANCIAL MOBILE SYSTEM BASED ON INTELLIGENT AGENTS in the Austrian Computer Society book series, Austria, 2006.
6. F. Zacaras Flores, R. Cuapa Canto, F. Lozano Torralba, A. Vzquez Flores, D. Zacarias Flores. u-Teacher: Ubiquitous learning approach, pp. 9–20, june 2008.
7. Alberto Tellez, Antonio Juarez, Gustavo Hernandez, Claudia Denicia, Esau Villatoro, Manuel Montes, Luis Villasenor, INAOE's Participation at QA@CLEF 2007, Laboratorio de Tecnologas del Lenguaje, Instituto Nacional de Astrofsica, ptica y Electrnica (INAOE), Mexico.
8. Izquierdo R., Ferrndez O., Ferrndez S., Toms D., Vicedo J.L., Martinez P. and Surez A. QALL-ME: Question Answering Learning technologies in a multiLingual and multiModal Envinroment, Departamento de Lenguajes y Sistemas Informticos, Universidad de Alicante.
9. http://java.sun.com/developer/technicalArticles/javaserverpages/wap
10. http://ilps.science.uva.nl/WikiXML/database.php
11. http://lucene.apache.org/

Semantic Routing for Structured Peer-to-Peer Networks

Luis Enrique Colmenares Guillén[1], Omar Ariosto Niño Prieto[2], Leandro Navarro Moldes[3]

[1] Benemerita Universidad Autonoma de Puebla,
Facultad de Ciencias de la Computación,
BUAP – FCC, Ciudad Universitaria,
Apartado Postal J-32,
Puebla, Pue. México.
lecolme@cs.buap.mx,

[2] Université Claude Bernard Lyon 1
Bâtiment Nautibus
43, Boulevard du 11 novembre 1918
69622 Villeurbanne Cedex France
OMAR.NINO-PRIETO@bvra.etu.univ-lyon1.fr

[3] Universidad Politecnica de Cataluña,
Jordi Girona, 1-3
Barcelona, España
leandro@ac.upc.edu

Abstract. During this work a simulator of a mechanism of searching in collaborative application is made to manage the content through the WWW with structured Peer-to-Peer networks. The principal contribution of this work is one mechanism that joins a metadata in the queries that are sent by the DHT Routing. The Semantic Routing proposed reduces the number of the average hops to reach an object or a document. This routing improves the Bamboo-DHT in the search of documents.

1 Introduction

The distributed computing introduces new challenges because it has new resources that show very different characteristics with heterogeneous architectures connected by distinct networks.

The increasing of Internet resources has been growth up, so the necessity to manage them has been initiated. This management has become an obligation of any distributed system that search the best functionality for future applications.

It is convenient to stand out that the negotiation of the resources consist to facilitate by the optimum way the available resources on Internet, databases, big stored systems or the files systems, among others. It is to be able to predict failures, network errors, traffic and other requirements.

The decentralization is important in applications where having global information is impossible. The decentralization gives to the networks or applications a high scalability and do not allow to have a single point of failure. The peers are autonomous and in some cases they are anonymous. The decentralization generates challenges in the security for different policies of management and dynamism.

The high dynamism of the participants requires new services. These services like the self-organization, replication, searching, among others. Those tasks give performance to the distributed applications in the context of collaboration group.

1.1 Content Management System

There are three important characteristics of the contents that can be used for the proposal of the management system with Peer-to-Peer (P2P) Networks:

1. The growing of the non structured data [1].
2. The management of the content of the network edge is considered an extension of the management of documents. The management of contents must be oriented to the reuse of the content.
3. The collaboration for Internet between the users of universities is frequent and useful in the actuality. The collaborative applications like BSCW [2], MOODLE [3] like others, are client-server in the WWW, the contents is in the server and the participants depend on the robustness of the server for having the active contents.

The idea of using P2P networks is not new in the management of contents. Several proposals exist for the management of the content in the companies [4] and some applications where the content is distributed by the audio, i.e. Gnutella [5].

2 The proposal requirements

The proposal is derived to offer solutions in the scenario that is in these three suggested dimensions, Fig 1:

1. The roles indicate the activity degree that the participants that have and their types of messages specifying protocols based on messages that allows guaranteeing a successful sending and reception of messages.
2. Distributes Hash Table (DHT): they are a weakly-connected_and strongly-connected this last, gives consistency the table routing. Today, the strongly-connected DHT have the lookup with O (log N) to O (log (1)).
3. During the earlier works, the data storage had been oriented from adjust of the system index to the file replication. The cache is emphasized to save temporally, during the restoration mechanisms of the routing tables until the defined roles like in JXTA [10].

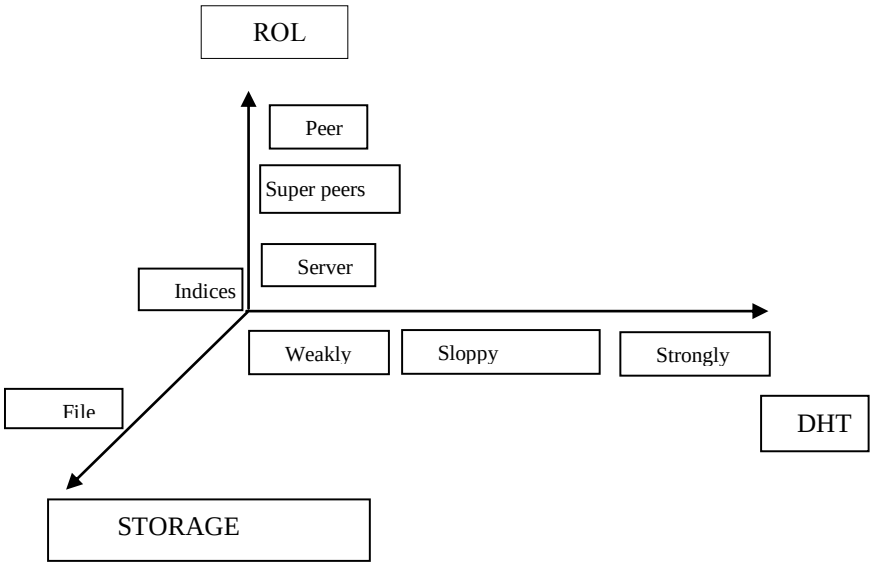


Fig. 1. The application dimensions

3 The Simulator proposal

In this simulator a brief description is made.

1. Classes and scripts with the configuration file. The design and the operation are explained.
2. The different events that the class DHT and semantic has.

3.1 Scripts and Classes with the file of configuration explaining the design and the running

In the programming guide, the design and style of the source code of bamboo [6] can be observed. Two stages or classes were created. The first class gives the number of hops through the DHT routing or the semantic routing. The second class is for locating the puts and generates the gets. The puts will place the objects (documents) through the bamboo nodes and the gets will find the request. Moreover, two scripts that have two classes.

The two Scripts are made in Perl. The first script is the responsible for creating the nodes type bamboo network, moreover of sending the java threads to the nodes of a cluster, balancing on each node in the cluster, the number of bamboo network nodes. For example, if there are 500 Bamboo-DHT nodes and we have ten nodes of a cluster (the nodes have dual microprocessors); we would have 50 bamboo nodes in each node

of the cluster having a 25/CPU relation. This script calls the *name.cfg* configuration file and this also calls the DHT class and semantic class.

In the article [9], the chum rate is made with ranks of 8/seconds to 4/minutes. This means that following the experiment, the paper shows that every 8 seconds 8 nodes are disconnected and 8 nodes are connected maintaining the session time average in 0.72 minutes with 500 nodes. The highest rank is about 4 nodes that are disconnected every minute and four nodes that are connected during the same minute maintaining a session time of 86.64 minutes.

Example:

$t_{med} = N \ln 2/\lambda$ where N = number of nodes, with *churn-rate* from 8/seconds to 4/minutes with one t_{med} =1.4 minutes to 3 hours. In the experiment of [9], they have N = 1000 nodes. The estimation ln(2)/ 0.008 = 86.6433976 = 1.44 minutes. In our case N = 500 nodes, then 8/500 =0.016, ln 2/0.016 = 43.3216 = 0.7220 minutes.

In the formula $\ln(2)/\lambda$. Where, λ= (4/60) / (1000) =0.00006666666. The total is = 10397.2181 = 173.28 minutes almost three hours. With N = 500, (4/60) / 500 = 0.00013333. ln(2)/0.0001333 = 5198.603867 = 86.643397 minutes.

The second script calls the class puts-gets that is located in the class DHT that is of bamboo. The puts follow the law of Zipf and they are made from parallel form in each node. The requests are made by each node that asks for objects with a balanced form. In other words, if there are 1000 requests and if there are 500 nodes, 20 requests are sent per node and each node will follow the Zipf law with random objects that the bamboo network has. Remember than the objects were put before by the puts class.

Using the DHT class calculates the number of hops using the Bamboo-DHT routing. Using the Semantic class, the number of hops is calculated using the semantic routing.

The limits can be changed, for example:

- The number of networks of the contents. It is relation with the distinct gateways of the beginning that will provide configuration file; this will generate different bamboo-network of content. This result that each bamboo network of content can make the next variations
    1. The number of nodes that learn: It is relation with the total number of bamboo nodes, we choose the random number and it will tell us the nodes which will learn, these change can be done with the class DHT or semantic.
    2. The number of objects: they are all the total documents that we can random put in the entire bamboo network; there the class puts/gets is found.
    3. Storage in each node: During the time that the requests are made, there exist a maximum number of objects that can store each node and it's found in the DHT semantic class.
    4. The requests following the Zipf law: The node that makes the request following the Zipf law. Making sure that the popularity of the objects is found in the puts/gets class.

Using the DHT class, calculates the number of hops using the Bamboo-DHT routing. Using the semantic class, the numbers of hops are calculated using the semantic routing.

Example of a configuration files in Bamboo, *name_file. cfg*. Each node of the bamboo network must have a fill of these characteristics

```
<sandstorm>
    <global>
    <initargs>
        node_id ${NodeID}
    </initargs>
    </global>

    <stages>

    <Network>
        class bamboo.lss.Network
        <initargs>
#               udpcc_debug_level 0
                drop_prob 0.1
        </initargs>
    </Network>

    <Rpc>
        class bamboo.lss.Rpc
        <initargs>
        </initargs>
    </Rpc>

    <Router>
        class bamboo.router.Router
        <initargs>
#         debug_level       1
          gateway                ${GatewayID}
          periodic_ping_period   20
          ls_alarm_period        4
          near_rt_alarm_period   0
          far_rt_alarm_period    10
#           leaf_set_size           8
          digit_values           2
          ignore_proximity     false
            location_cache_size    0
        </initargs>
    </Router>

    <DataManager>
```

```
        class bamboo.dmgr.DataManager
        <initargs>
          debug_level           0
                merkle_tree_expansion 2
               desired_replicas       2
        </initargs>
    </DataManager>

    <StorageManager>
        class bamboo.db.StorageManager
        <initargs>
#         debug_level             0
          homedir       ${CacheDir}
        </initargs>
    </StorageManager>

    <DataManagerTest>
        class bamboo.dmgr.DataManagerTest
        <initargs>
 #               debug_level      0
                to_put           0
                put_size       100
        </initargs>
    </DataManagerTest>

        <Dht>
           class bamboo.dht.Dht
            <initargs>
                #debug_level 1
                storage_manager_stage StorageManager
                min_replica_count     1
            </initargs>
        </Dht>

        <Gateway>
           class bamboo.dht.Gateway
            <initargs>
               # debug_level 1
                port ${GatewayPort}
            </initargs>
        </Gateway>

    <WebInterface>
        class bamboo.www.WebInterface
        <initargs>
                storage_manager_stage StorageManager
        </initargs>
    </WebInterface>
```

```
#Routing dht_bamboo
    <StageSemantic_dht>
        class bamboo.cont_dht.StageSemantic_dht
        <initargs>
          debug_level 1
        </initargs>
    </StageSemantic_dht>

    </stages>
</sandstorm>
```

The example of the *name_file.cfg*, it is a structure in the following way: begins with < sandstorm > and concludes with </sandstorm>. In the global parameters the *node_id* contains the IP of the participant; it in this case, puts the IP of the node of the cluster and the port that distinguish each node. Stages that are bamboo classes are contained. Each stage contains parameters that allow controlling certain characteristics of a physical network. Certain stages were used to allow the creation of a bamboo network like in the reference [7]. The stage *Network* generates parameters of a network and the stage *Rpc* allows the remote procedure call. Router allows using the routes based on the bamboo-DHT considering two groups of neighbors (leaf-Set and table Routing). *DataManager* manages the number of replicas that can store when we use a Put. *StorageManager* allows keeping the data physically in the directory that is indicated. DHT calls to *StorageManager*. Gateway indicates that port used each node for the communication and the step of messages. The stage *WebInterface* allows to make call with the protocol *http* in a navigator to show the bamboo nodes. We conclude with *StageSemantic_dht*, this class allows using our algorithm which adds additional caching to each participant and allows choosing other possible routing.

For example: for a node that initializes the cluster that has IP, 196.0.0.1 the first node, the second 196.0.0.2 and so forth until the node 30 of the cluster, 196.0.0.30. If begins with the port number 3630. The first node 196.0.0.1:3630. The port 3631 reserves for the step of messages (gateway_port of the stage gateway of the file name_file.cfg) and the port 3632 is used to visualize in a graphic window the bamboo nodes. In ten bamboo nodes in a cluster node, would be made increasing of three in three each port and the address IP. Example: 196.0.0.1:3633, 196.0.0.1:3636... 196.0.0.1:3657.

The entire DHT networks have an origin. Bamboo uses the gateway of the stage Router, i.e. all the nodes belong to the bamboo network that is formed with the same gateway. In this case, has the port beginning of the first node 3630. If the bamboo network begins in the nodo1 of the cluster, all the nodes of our application have the following node gateway: 196.0.0.1:3630.

In the stage *StorageManager* to have a directory to locate the components of an object physically: value and content. Moreover, the directory /tmp is used.

This simulator allows that can be used the class java to scenarios with real nodes. It is transparent to the user and can expand the work to nodes PlanetLab [8]. Now, the open-DHT and bamboo are in the nodes PlanetLab.

The simulator works as follows: Once that execute the first script create nodes Bamboo-DHT network. The second script put to the objects in the all bamboo network. The objects are distributed in a uniform way in the bamboo network. The following step is send the class Gets and to do the requests of the objects. For example: suppose that are 100 objects and 50 nodes bamboo, a node requests an object for the first time, the algorithm calculates the number of hops and sends a message to make the copy to the node in its caching and this guarantees that if any node sends a requests the object and if during the journey have the object then has it reduces the number of hops. If makes the request again, now reduce in smaller number of hops because now already has the object.

3.2 The events that to have the class DHT or semantic

The events that use the DHT Routing (DR) or semantic Routing (SR) are:

1. Types of messages.
2. The payload or what is the message.
3. The different logs files. Format of Logs. The logs stay in each node bamboo:
    - Log of requests that indicates the objects that stay in the node.
    - Log of number of hops to reach an object. It registers the object number and the number of hops that it required to arrive to the object.
    - Log of the DHT, is when the bamboo-DHT is used to reach the object.
    - Log of average hops of an object, the maximum of hops of an object, minimum of hops of an object, and number of times of the object.
    - Log that counts number of total objects for each object. In this log, helps identify how many times they are requested each object, remember, that they are objects that are located in the bamboo network.
    - Log that counts the numbers of total hops.

4 Evaluation

4.1 Platform

The platform to measure our architecture this made in Bamboo-DHT [9] and the language scripting is Perl; the code is within a cluster of 16 nodes. The nodes of cluster have dual microprocessors AMD Opteron 64, 1.4 GHz, connected by Ethernet Gigabit, RAM memory of 2 GB a hard disk SCSI of 36 GB and OS fedora Linux Core 6x86-64. The simulations are execution-driven. All the file logs are stored in each participant of the network overlay P2P in run time.

4.2 Evaluation

For the evaluation three modules or classes in Bamboo were added. The first module, is caching, that assigns cache in each participant-DHT and allows the node to choose between two routings. The second module makes simultaneous requests in parallel. This allows making requests of queries from different participant nodes at the same time. The third module is the replication of content. This module will consist basically of the extension of the epidemic algorithm of Bamboo-DHT to update the contents in caching of each node.

4.3 Static Scenes

The evaluations are made both routing SR and DR. In Table 1, shows the parameter that was used for static scenes with Bamboo-DHT routing. The number of tests that were made was approximately 100 by each scene, although from test 45 the graphical one shows a uniform behavior with respect to the growth of the average of the number of hops.

The Table 2 shows the parameter that was used for static scenes with SR. A metadata in query of DHT is used; in addition caching that was made vary in the size of each node or participant. The number of objects is important, if are a small number of requests, 5, 10 or 20 requests. The objects will vary of 100, 250, 500, 750, and 1000 with duration of the object of one week (7*24*3600) or a day (24*3600) guaranteeing that the objects always are in the bamboo network. The communications between nodes of the cluster are stable in approximately 85% of the nodes. This guarantees 85 % of objects. In the number of requests: 10000, 12500, 15000, 20000, observed that the limits of the average are reduced. The size of caching used are: 5, 10, 15, 20. The use of caching represents use of DHT in SR. In scenes 1, 2, 3 of table 2, using small caching, would use of Bamboo-DHT. In experiments 4, 5, 6 used greater caching, would use less the Bamboo-DHT that is reflected in a smaller number of hops.

Table 1. Bamboo-DHT Routing parameters.

Scenes: Bamboo-DHT	No. Nodes (parameter 2)	No. Objects	No. Requests (parameter 1)	Without *caching*, without metadata
1	500	100	10000	----------
2	500	250	12500	----------
3	500	500	15000	----------
4	500	750	15000	----------
5	500	1000	20000	----------

Table 2. Semantic Routing parameters.

Scenes: info-semantic (parameter 4)	Nodes (parameter 2)	Objects	Requests (parameter 1)	Size of caching (parameter 3)
1	500	100	10000	5
2	500	250	12500	5
3	500	500	15000	5
4	500	750	15000	10
5	500	1000	20000	15
6	500	1000	20000	20

In Fig. 2, the minimum, 3er quartile and the average of the number of hops by searching each object were found. The graph of Fig. 2 shows the following: The five more popular average objects, the total of objects represent 95% of the total requests. The requests are sent from 500 nodes of a total of 500 available nodes. Each node sends *n* requests of *m* available object. Each experiment of 500 nodes uses 500**n* requests. In *x*-axis, the objects are ordered from the most popular to the least popular, indicated by the named percentage and with their labels from left to right. In the *y*-axis, the number of total hops is shown.

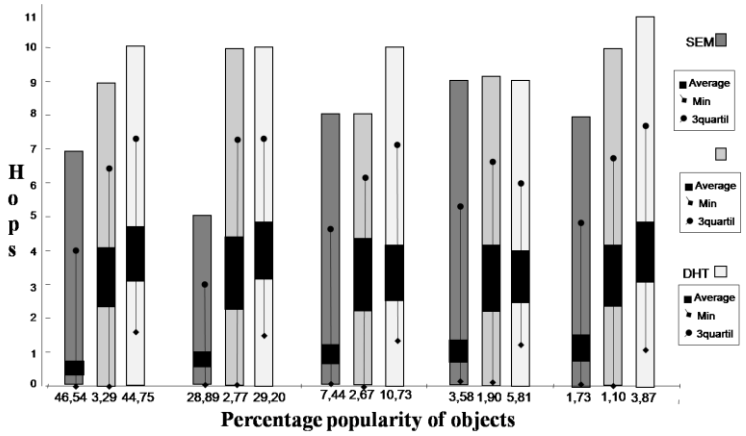


Fig. 2. Comparison of Routings.

In Fig. 2, there are five most popular objects with three bars, respectively. The two first bars of each object represent semantic routing proposed and it is a combination of both routings. The last bar of each object represents Bamboo-DHT routing without alterations. The bars are divided in quartiles; the maximum number of hops for a same object is the high part of the bar. The third quartile represents the superior part of the line that is within the bar, the average or second quartile represents the high part of the black picture and the first quartile is the minimum hop that was obtained by requests of a same object that represents the final part of the line. The average of hops diminishes when the object is more popular. We observed in the graph that improvement in the number of hops by DR and SR are the two initial bars. DR represents the third bar of each popular object; the average is over the two bars.

5 Conclusions and Future Work

The main contribution of this work is to improve the routing of the Bamboo-DHT, with semantic routing that reduces the number of hops average. The culmination of this work will have two contributions, first: 1. a mechanism of content delivery that allows possible routings to reach the content, when adding semantic routing in Bamboo-DHT, obtained a reduction in the number of hops averages to find an object and 2. A search mechanism using semantic information, giving expressivity to DHT, single-key was used and it does not affect the costs of the communication between nodes of DHT.

In future work, the last contribution will end with the class of bamboo that extends the broadcast algorithm, in addition, the evaluation in dynamic scenes, with parameters likes churn-rate and limitation of resources.

References

1. Robert Blumberg, Shaku Atre. The Problem with Unstructured Data. DM Review Magazine, February 2003.
2. BSCW (Basic Support for Cooperative Work) http://bscw.fit.fraunhofer.de/ (2009).
3. *Mo*dular Object-Oriented Dynamic Learning Environment (Moodle) http://moodle.org (2009).
4. “The emergence of distributed content management and Peer-to-Peer Content Networks”, by Gardner Consulting, January 2001.
5. T. Klinberg and R. Manfredi, http://www.gnutella.com/, 2009.
6. http://www.bamboo-dht.org/programmers-guide.html, 2009.
7. http://www.bamboo-dht.org/, 2009.
8. http://www.planet-lab.org/, 2009.
9. S. Rhea, D. Geels, T. Roscoe and J. Kubiatowicz. Handling churn in a DHT. In Proc. of the use USENIX annual technical conference, June 2004.
10. Juxtapose, JXTA. https://jxta.dev.java.net/ 2009.

A Recommender Agent Development

Juan C. Ramirez[1], Darnes Vilariño[1], Fabiola López[2],

[1] Faculty of Computer Science, Benemerita Univesiad Autonoma de Puebla (BUAP), Av. San Claudio s/n esq. 14 sur, Ciudad Universitaria, Puebla, México.
[2] Direccion General de Innovacion Educativa, Benemerita Univesiad Autonoma de Puebla (BUAP), Av. San Claudio s/n esq. 22 sur, Ciudad Universitaria, Puebla, México.
jcramirezmx@gmail.com, darnes@cs.buap.mx, fabiola.lopez@siu.buap.mx

Abstract. We present the development of different modules from a recommender agent, an update of the algorithm for rating services; and besides, some learning rules needed to properly recommend services. Recommender agents have been a great help for users in Web's searching processes and information retrieval. This paper's proposal to show the development advances in a recommender agent who links agent theory, recommender systems, and virtual worlds, which becomes a highly innovative subject. Programming of diverse agent modules has been developed on Linden Language Script (LSL), Visual C#, and MySQL. Currently we have a prototype in our server.

Keywords: Recommender Agent, user profile, learning rules, services.

1 Introduction

Nowadays education is leaving behind traditional teaching methods and adopting new learning techniques thus, new concepts arise as e-Learning, virtual learning environments, among others. E-Learning is the educational process placed through mechanisms supported by Internet technologies using audio, video, text, and graphics transmissions [1].

There are places around the world where learning techniques based on virtual worlds are highly used. Teachers and students coexisting in the same virtual environment (like a classroom), but all of them in different locations is a very interesting idea, so many projects are being developed to implement this teaching technique in near future.

There are several approaches trying to combine virtual worlds environments and e-learning systems, an example of this is Sloodle [2], which is a fusion of Secondlife environment and Moodle platform.

Moreover, Retrieval Information area has taken great boom among researchers by the need to develop new software techniques for helping users in search processes. Usually a user loses a lot of time finding interesting information looking for quantity or quality of it.

A solution for this problem is using recommender agents. The function of these agents is to search information related to a user and, as the name involves, recommend to him.

Recommender Systems try to be one step forward from traditional information retrieve, which works by using keywords to find information about a topic through the well-known search engines (google, lycos, yahoo, etc). As the name indicates the recommender systems recommend or suggest to the users the items or products based on their preferences, they are used by Web sites, like electronic commerce, as marketing tools in order to increase the sells presenting to the user those products that user wishes (or may wish) to buy. Thus, we can know what customers like or dislike [3].

Nowadays many websites offer recommendations to their users, these sites are mainly focused to online sells, and of course, the recommendation systems are focused to that area too. An example of this is the E-cousal system, which is based on catalogue sells and supported by a web page and an application functioning as a multi-agent system. [4].

In virtual worlds, users are represented by avatars and they can interact among each other in a 3D world. In those environments there are different services offered to them, with a drawback: the larger a virtual world becomes, the harder it is to find new services or activities added to it (just like the real world: the larger a city becomes, the harder it is to find certain places), these services could increase in a chaotic and uncontrolled way, causing serious limitations for their management, organization and retrieving.

The paper's proposal is to show the advances in the developing of an agent in a virtual world; showing how is built the user profile and the implementation of different modules of the system.

In Section 2 we describe architecture of the agents; in Section 3 we talk about the services and how they are managed; in Section 4 we show the implementation of 4 modules: database, the user profile, the algorithm for rating recommended services and some learning rules; in section 5 we describe some results and finally in section 6 the conclusions and future work.

2 Architecture

The main components of the recommender agent are shown in Figure 1 [5]. Components are listed on rectangles whereas the exchange of information is listed with arrows unidirectional or bidirectional depending on the function. All of these run under OpenSim platform, which is an open source virtual world simulator [6]. In this work we explain in detail how we build the user profile, the Database implementation, and part of the learning techniques, an improvement of the algorithm for rating services and how the services used by the recommender agent are managed.

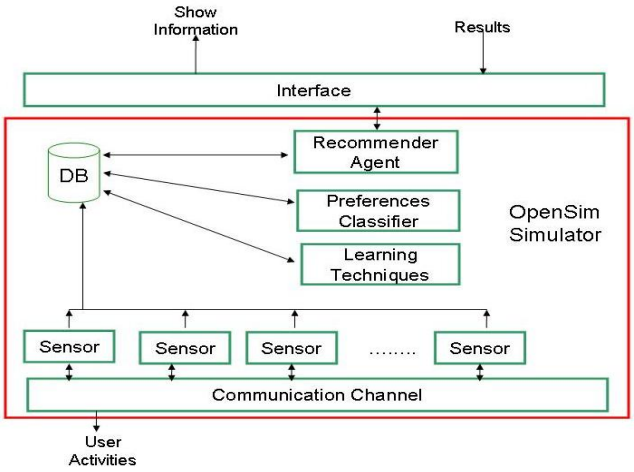


Fig 1. System Architecture

3 Services Module

The services in a virtual world could be varied, from recreational activities to debates or discussions about a specific topic. Services are divided into these categories: Introductory courses, Conferences and Interactive talks, Classes, School courses, Recreational courses, and Parties, luncheons and events.

At this moment the creator of a service is responsible to introduce the service information in the database, filling the required information in a Web page. The data required are: service description, title, keywords and type of service among others. The recommender agent needs this information in order to recommend services. The options for the service creator are shown in Figure 2.

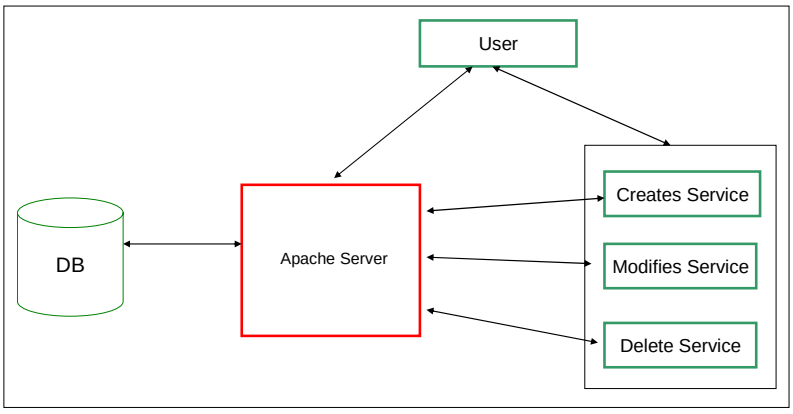


Fig 2. Services Registration Architecture

The user access to their services through a web page bases on Apache server, it allows to create, modify, or delete a service related to that user. When a user creates a service the data required is inserted as a new row in the database and it's related to the user, if the user modifies a service, previously created, the information is updated on the database. The user also can delete a service, in this case the row is not deleted from the database, and the service is only marked as disable.

Using this page, users can offer their services and they can be recommended to other users. Also this page allows the management of services related to a user.

4 Implementation

4.1 Database Implementation

OpenSim simulator, mentioned above, includes a database which is essential for simulator's proper working, and in which users, objects, regions and scripts' registration take place in the virtual world. The database has a table named users, in which user IDs (UUIDs) are defined, those IDs are used to link original database tables with new tables needed by the recommender agent in order to meet their goals.

Figure 3 shows how the tables are linked to the existent users table, in one hand it's linked to the user profile table, which in turn is connected to the topics that the user prefers. In the same way, the activities are related to a user by his ID. Finally we store the recommended services' list for subsequent rating.

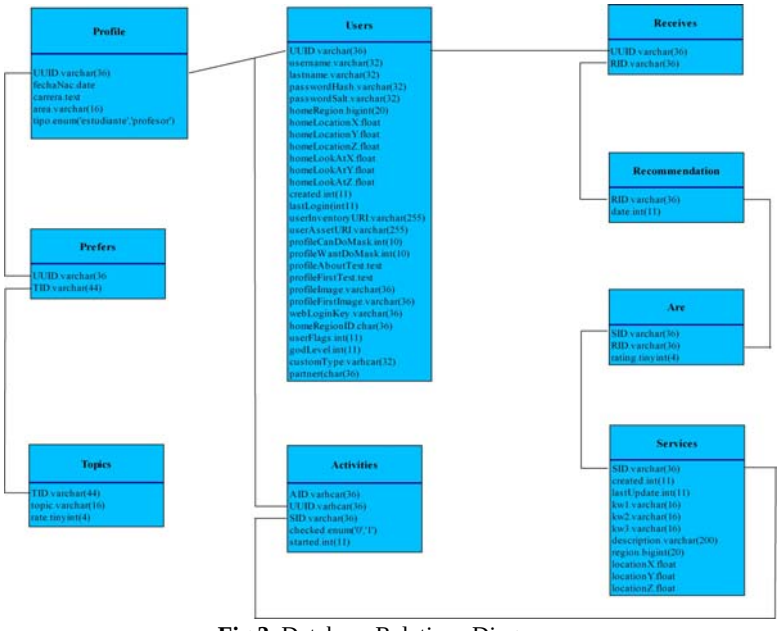


Fig 3. Database Relations Diagram.

Table *Profile* stores the user information, it has a foreign key UUID (users), that is because profile is just an extension of the user information, but the table users is already defined in the original Opensim database and we didn't want to modify those tables.

Table *Topics* store the keywords associated to services, these keywords are the themes that user prefers. This table has its primary key TID, the name of the topic and its rate.

Table *Prefers* connect the table profile and the table topics, it has a primary key which references to UUID (profile) and TID (topics). This way we can now what topics a user prefers.

Table *Services* stores the services created by users and registered in the services web page, its primary key is SID, and has more information as created, description, region, in order to know where the object is.

Table *Activities* registers any interaction between the user and an object (which can be linked to a service), it has a primary key AID, and two foreign keys, UUID (users) and SID (services). These activities are not stored permanently in the database, once the activity has been analyzed, it is marked as "checked" and then deleted.

Table *Recommendation* stores the recommended services list sent to the users, only has its RID and the date the list was sent.

Table *Are* is connected to table Recommendations and Services, it connects the services with a recommendation list. Its primary key references to SID (services) and RID (recommendation).

The last table is Receives, which connects a recommendation list to a user Its primary key references to UUID (users) and RID (recommendation).

4.2 Building user profile

Sensors are objects located along the virtual world, by which they listen the communication channel used in OpenSim (registers interaction between avatars and objects associated to services in simulator). These sensors are programmed using Linden Script Language and the information retrieved by them is sent to a web page in order to store that data in the database, this is because the language is very simple and it doesn't allow connection to databases or output of information into a file. Sending the data to a web page is very helpful, and using PHP (which is a robust language) we can connect to databases.

Sensors register information in table activities in the database, ensuring availability for future analysis. This table is continuously handled by the preferences classifier agent to debug information stored in it. The agent reads this table, and according to the information read, inserts or increases the rating in table topics. After doing this, the agent delete the information previously read in order to avoid garbage data stored in the database.

In figure 4, we show the interaction of sensors, the preferences classifier agent and the tables in the database. This is part of the architecture, but with more details showing the process to build the user profile.

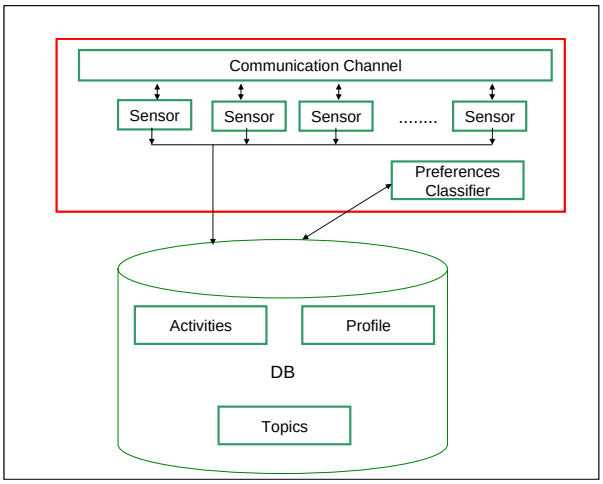


Fig 4. Building a User Profile Architecture

Finally it is worth noting that the sensors inside the virtual world are in a passive form, they only changes to an active form when an avatar is close to them, and then, they starts to collect the information.

4.3 Algorithm for rating recommended services

Tasks related to the preferences classifier agent are divided in two phases: Phase 1 Activities analysis; Phase 2 Rating analysis [5].

In phase 1, the agent continuously monitors database, where services, which avatar has interacted with, are stored, it classifies the services by preference (p) and by length (l). Preference is measured by services accessed frequently by the avatar, and length is measured by the elapsed time since the avatar started to use the service until it finishes using it. The agent applies Algorithm 2 in this phase and Algorithm 3 on phase 2 [5].

We propose new values for the ratings of recommended services in the Algorithm 3. With previous values the rating just increase its value showing only positive preferences, nevertheless using that values we can't know if the user dislike that service or it just doesn't matter to him. Moreover a problem arises just by increasing the rating: "is there a bound for the rating value or could it increase without a limit?"

The previous values were: increase 0.1 if the user rating is 1, increase 0.2 if the user rating is 2, same for rating on 3, 4 and 5, if user rates 0 do nothing. Below we show the algorithm with the new values:

Algorithm 3
Step 1: The agent retrieves services rated by the user in table recommendations
Step 2: Select keyword from table services in order to update or insert it in table topics
Step 3: If rating associated to service is:
0: Decreases rating in 0.3

1: Decreases rating in 0.2
2: Decreases rating in 0.1
3: Increases rating in 0.1
4: Increases rating in 0.2
5: Increases rating in 0.3

The values description is explained in the table 1.

Table 1. Description of the values

Value	Description
0	The service dislikes very much
1	The service dislikes
2	The services dislikes a little
3	The services likes a little
4	The service likes
5	The service likes very much

We can note that, according to the rules for rating services, the rating number can be a negative value. This way we can now know that a user doesn't like certain topic.

4.4 Learning rules

A first approach to learning rules has been developed. The rules try to solve the posterity problem; this is for topics that the user rates as a like a time ago, but now that topic has no more interest for the user. For example: a year ago a user interacted with topics about Olympic games, and that topic has a rate of 1.5, now the user is no more interested in Olympic games, but the recommender agent continues recommending services related to that topic because of its rating. There is not enough (and also not advisable) that the user rate several times the service recommended with 0, in order to decrease its value. We can apply some rules that, given a period of time, analyses if the user has been interested in a topic. If the topic has not been modified in that period of time we can "reset" the value of the topic setting it on a value of 0.

```
If a topic rating has not been modified in X period of time
Then     if its value is negative
         Then do nothing
         Else reset its value to 0.
Else     //a topic rating has been modified in X period of time
         if it has been modified only decreasing it
         Then reset its value to 0.
         Else do nothing.
```

These rules are very simple but solve the posterity problem explained above. More rules will be developed for working analyzing user activities and updating the user profile

5 Results

Web page access to management of services is working properly and there has not had problems. The page is based on Apache server using PHP and MySQL. Services related to users has been added and stored in the data base, and the modification or deletion of them works fine too. Relations of the tables in the database fulfill its role, and no problems have been reported related to it.

Sensors register information read on the communication channel and we are doing some tests to ensure that the web page using PHP stores the data efficiently in the database, one disadvantage of using web page is that the server may be saturated with connections, but at the moment this is the best solution.

The Learning Techniques module is under construction and the learning rules have not been implemented yet, but we hope shortly have some results to show.

6 Conclusions and Future Work

This paper shows an important advance in the implementation of the recommender agent, step by step the different modules are been developed, and they works properly. There has had some complications developing the sensors but they has been successfully solved bypassing the language and the database through a PHP webpage.

Currently the user profile is being developed and analyzing it in order to find the optimal balance between the user preferences and user activities. Agent learning capacities are based on deductive rules, which are being created. Only the rules presented in this work have been developed, but when all the rules are ready, the agent could infer the best services to recommend to the user.

References

1. Marcus, A. “Caracterización de los Servicios de la Educación a Distancia desde la óptica de una plataforma distribuida orientada a objetos”. Master degree thesis ITESM. Monterrey, México., 2000.
2. Daniel Livingstone, Judith Kemp, and Paul Andrews. Sloodle Learning System for Virtual Environments. http://www.sloodle.org, 2009. Accessed on February 23, 2009.
3. Oswaldo Velez-Langs, Carlos Santos. “Sistemas Recomendadores: Un enfoque desde los algoritmos genéticos”,. Industrial Data, ISSN: 15609146, Volumen 9, Fascículo 1, Pag. 20 - 27.

4. Ana Gil, Zahia Guessoum, Francisco García. “Recomendadores en un Sistema Multiagente Adaptativo para el Comercio Electrónico”,. Taller en Sistemas Hipermedia Colaborativos y Adaptativos dentro de las JISBD’2002. El Escorial, 18 al 22 de Noviembre del 2002.
5. Ramirez, J.C., Vilariño, D., López F., (2009). A Recommender Agent Design For Virtual Worlds. Proceedings of the IASK International Conference – E-Activity and Leading Technologies. pp 142-146. ISBN: 978-989-95806-7-1.
6. OpenSimulator Web page. http://opensimulator.org

Some Considerations for the Semantic Web

María Elena Franco Carcedo

ICUAP, CA RI-FCC-BUAP, Ciudad Universitaria,
72570 Puebla, Mexico
fcarcedo@hotmail.com

Abstract. Here are some considerations about the management and level of verbal language domain of semantic web potential users, their characteristics, most frequent errors and consequences, in order to be took into account for a better performance of the semantic Web.

Keywords: Semantic Web, verbal language.

1 Introduction

This paper has been constructed on the basis of vast experience gained through several years on the characterization of the academic language (record/register). Moreover, this research is the result of a deeper study on postgraduate careers of a public university (BUAP), which let us to detect the most frequent mistakes and transgressions on the academic language.

We will start this discussion with two questions: 1) What is the meaning of those terms that we hear and use, but whose conceptualization or definition will state difficulties of some kind? In academia, especially in recent years, there are some terms that are frequently used within institutional development plans. Terms such as: pedagogy, didactics, methodology, teaching techniques, teaching process learning, etc are used as a prelude of academic excellence and resource optimization (human resources and materials). And here becomes the second question: Does not imply knowledge any kid of specific knowledge, a specific language, a certain jargon? If we accept this assertion as a claim we may infer that by acquiring a scientific knowledge[1] of language (knowledge instrument[2]) we can facilitate the language knowledge and application of the same. But the real question is the following: Does the use of a

1 There is a theoretical discussion about the status of the studies dealing with the language, from *grammar* to the diverse branches, lines and research. By centuries it was considered the *grammar* and *rhetoric* and *oratory* as complete arts. However for the end of last century, with F. de Saussure as a benchmark, it was understood as a *science*. In this paper we will no go deeper in order to argue about one or other viewpoint.

2 We would argue whether or not the natural language is the knowledge instrument or the knowledge itself with respect to its role as reality creator, in terms of the *rational speech* or thought. *How do we know*? *How do we learn the reality*? By means of which mechanism or procedure we "know" and "know the world".

semantic Web require some user verbal skills or at least to advise the lack of these skills?

2 Considerations

Errors or deficiencies affect at different levels the following issues: *intelligibility* (Rules 1, 2, 3 of verbal system); *the thought logic* (Rules 3, 4, 5); *communication* (null or minimal noise in the process of encoding-decoding); ***speech*** *fluidity and precision* (speech alludes to both phrasing and reasoning). Natural language is a system of verbal symbols (sorted set of functioning rules of constituent items). The academic subsystem (record/register) and jargon (characteristic of the different areas of knowledge) is characterized by the following rules.

2.1 Phonetic-phonological and spelling rules

- Signs Graphics: punctuation and auxiliary signs.
- Spelling: diacritics, punctuation, accentuation, ortho-semantic problems.

2.2 Lexical-semantics rules

Lexical (3rd unit of language, 1st significant unit): univocity, precision, accuracy, conciseness, versatility and richness vs. ambiguity, categorization error, polysemy, synonymy, homonymy, homophony, paronymie, imprecision, vagueness, semantic phrases, poverty, repetition, ignorance, confusion.

2.3 Morphosyntactic rules

- Constructive (5th unit of language, 1st logical unit): logic, consistency, sequential, cohesion, consistency, clarity, simplicity vs. illogical (juxtaposed and copulative; alteration of conjunctive nexus, etc.) assumptions, fractures, anacolutha, truncated sentences, inconsistency (especially subject-predicate), confusion (especially in terms of managing the elements of the compound sentence) , scavenging, and Pseudo-cultism (ex, abuse of “el cual” - which); incorrect temporal correlations, and so on.
- Concordance, among others.

2.4 Encoding rules

- Encoding: denotative, which corresponds to the formal record/register (academic, scientific, technical) vs. subsystem of use, characterized by a

tendency of connotative encoding, the standard or colloquial register, oral, with semantic shifts or meanings tagged by chrono- and geo- sociolects –the use of archaisms (a word or form of speech no longer in common use); or only functional in some areas or among certain social groups – altered prepositional regimen, etc.

2.5 Logical-stylistic rules

- Structure (syntax of the compound sentence, what relationships are established: cause-consequence, principle-purpose, conditional potential, hypothesis, etc.); the ideas exposition and their linking; structure consistent with the basic genre on introduction (definition, description, narrative, argumentation), development and conclusion.
- Functional: the event affects the purpose, theme (or its treatment), target and record/register (what, for whom, for what purpose, how); informative function, logic, data or knowledge transfer, assertive or speculative, formalized as article, essay, notes, summary; without subjective, intimate or emotional levels.
- Style: neutral and objective tone; author plural or third impersonal; without dialect marks; rigorous, light, clear and simple.

We are particularly interested on those problems related with the *word* unit. Therefore, we will not stress on unacceptable underlying assumptions, stemmed phrases, anacolutha and diverse fractures, erroneous punctuation, with the resulting disruption of the correct decoding or in cases of disagreement, but we will focus our attention on aspects related with Rules 1 and 2: **Phonetic-phonological and spelling rules**; **Lexical-semantics rules: transgressions and most frequent errors** and we will present some examples.

- Vocabulary deficiencies and use meanings[3]
- Confusion/alteration of logical-grammatical category
- Spelling errors that affect meaning (written accent mark: *íntegro-integro-integró; gráfica-grafica*, etc.).
- Problems with homophones, homonyms, paronymie and polysemy[4]
- Abuse of some Spanish verbs when defining and writing, such as: *ser*(being), *estar*(to be), *tener*(to have), *poder*(to be able), *hacer*(to do).
- False friends from the English language (Due to the users read literature in English: *remover* (remove) instead of *eliminar*, *interface* or *interfase* (interface) instead of *interfaz*, etc.).
- Abuse of lexical and constructive Anglicism and the problems they cause

3 In order to simplify and due to we are interested on the determination of the dialectic phenomena, we create the concept **of use** sub-system. We will focus on transgressions with respect to the academia norms that rule this record/register.

4 In both senses, strict and soft, i.e., we include here the meaning plurality.

- Misuse of the gerund (English gerund)
- Forms of 'passive English' (construction of the auxiliary 'to be' (*ser*) rather than reflected passive)

3 Experimental results

In this section we present an evaluation of a written test carried out on 135 students of different postgraduate programs on science and technology. We asked them to define different words. Even if the complete list was conformed by 25 terms, we only show the obtained results with a subset of this set. In summary, we have detected 13 problems, which are described as follows.

1. **Ignorance of the term meaning (null answer)**. Surprisingly, not one word were answered in blank (null answer), even on terms such as *vaso*(glass), *sección*(section), *adolescente*(teenager), where answer percentage rates were 1,48%, 2,22% and 9,62%, respectively. Percentages rates reach values such as 67,4% (*acerbo*), 65,92% (*asechar*) and 50,37% (*sito*).

- The pair *acerbo-acervo*, had 91 and 51 null answers (67,4% and 37,7% respectively), from which only 2 and 46 where partially correct (2,2% and 34%).
- The pair *adoleciente-adolescente*, had 34 and 13 null answers (25,1% and 9,6%), from which 37 and 69 were correct (from this 37, only 5 were totally correct, i.e., those that gave more than one meaning; in the second case, i.e., *adolescente*, there is only one meaning which may be expressed in different manners).
- The pair *baso-bazo*, had 66 and 34 null answers (44,8% and 25,1%), from which 29 and 93 where correct (partially correct in general, since in major cases there were given only one meaning) with percentages of 21,4% and 68,8%, respectively.

2. Confusion of logical-grammatical category within the term meaning.

1. Simple.

- **Term:** *soluble* (adj.) **Definition given:** 'disolver' (verb);
- **Term:** *cesión* (noun) **Definition given:** 'que se termina o *sesa [sic];*
- **Term:** *suspendido* **Definition given:** *'cuando* algo a [sic] terminado *ó [sic]* concluido'

2. Half confusion (one category inside and one category outside).

- **Term:** *acceso* (sust.) **Definition given:** *entrada* (sust.), *viable* (adj.);
- **Term:** *sesión* (sust.) **Definition given:** 'congregarse personas (verbo), reunión (sust.)'

3. Confusion with its homophones (diacritical spelling) or paronymie. The value ranges from 0% for *vaso* (*baso-bazo*) to 17,7%, 17,3% and 14,8% for *asechar, adoleciente, rebelar*, which are defined by their pair *acechar, adolescente, revelar*, respectively.
1. Homophone category error. *acerbo* (adj.)*: conjunto de bienes comunes* (corresponde a *acervo*, sust.)
2. Outside of its category [Example. *Acerbo* (adj.): *tener conocimientos* (they mislead with *acervo*, sust., but they defined it as verb)
3. Half confusion (one category inside and one category outside). *acerbo:* 'acumular, conjunto de cosas'
4. **Answer outside of the semantic field.** Unbelievable at firs glance, among the highest percentages it was found the term *vaya* (exclamation and verb) with the 77,77%, followed by *haciendo*, with 69,62%, *vario*, with 57,77%. Examples of wrong field:
- **Term:** *adolescente* [person] **Definition given:** 'etapa'; 'edad' [cosa];
- **Term:** *asciendo* (present tense verb, act) **Definition given:** 'posición en que uno queda' [resultado de un acto];
- **Term:** *revelar*, **Definition given:** 'acción *quimica* [sic] de la fotografía'
5. **Definition by means of**:
- Temporality: *es cuando* (ej. *estática*: cuando **una persona** no se mueve)
- Mode: *es como...*(it is like…) (ex. *occisa: es como estar muerta*)
- Location: *es donde...* (is where…) (ex. *jardín: donde juegan los niños*)
- Function: *sirve para...* (it is to…) (ex. *lenguaje: sirve para comunicarse*)
6. **To classify instead of defining**. Very often they use to annotate the infinite without definition (nor meaning) in the particular case of conjugated verbs. Example. *izo:* verbo izar; *vacilo*, 'de vacilar'.
7. **To complete instead of defining**. (example: *acervo: bibliográfico; sucesión: presidencial*)
8. **The efinition was to much…**
- Extended, wide (ex. *es una cosa...; es algo...*)
- Restricted (ex. *acervo: conjunto de palabras)*
- Fuzzy, imprecise (ex. *bacilo: pertenece a la biología*)
9. **Meanings or homonymy omisión** (we are only interested on those answers that show significant plurality but not the classification of the term meaning of homonym term). This is considered "partial correct answer". In *rebelar*, it was obtained a 41,1%; in *revelar* the 46,6%; in *ascender* the 48,8%, i.e., the majority.
10. **Figurative sense, without the denotative**. *árido*: 'falto de amenidad'
11. **Semantic shift**. *árido*: 'molesto'
12. **Subsystem of use or colloquial**. *vacilar*: 'echar relajo'
13. **Answer without meaning**. *arriar:* 'arroyar' [por *arrollar*]

In summary, below we show some other examples without classifying the specific error type.

- *Acervo*:: 'vacteria' [*sic*]; 'es la consistencia de algunas cosas, p.e. la piña'; 'afirmar'; 'contar algo'; 'hacer muchas cosas'
- *Revelar*: 'identificar caracteres'; 'presentar algo a escala'; 'quitar'; 'relativo a copiar, algo oculto'; 'descifrar'.
- *Vario*: 'del verbo *vasar [sic]'; 'alguna cosa de uso común'; 'mucho'*
- *Vacilo* 'que no tiene seguridad'; 'indeciso'

The above presented examples suggest taking some considerations in order to avoid that the requested information cannot be retrieved due to this type of user language fails. We do not expect an information retrieval system tries to "guess" what the user wanted to say from his written query (what he really said).

In the digital version of the Spanish Language Dictionary (Diccionario de la Lengua Española, RAE) there exist a simple mechanism of word 'approximation', i.e., if someone writes a word incorrectly (ex. 'grafica'), it suggest the correct word ('gráfica' or 'graficar'), i.e., it looks for orthographical similar words, which we consider to be a good example of semantic Web.

As conclusions, let us consider the following issues:

a) The language, any type of language, is a *symbol system*. A system is an ordered set of functional rules of the system items. In this case, we mean as symbols as *those that refer to something, which in general is different of itself*[5] *and with intentionality*[6].
b) The classical analysis[7] of the verbal sign demonstrate the use of three elements: the *phonetic-acoustic* or significant (and its graphical equivalence, i.e., the letters); the *eidetic-conceptual* or meaning, and the *Referent* or reality, i.e., the target object or what it is referred (it means, it denotes).
c) The *concept*, *the idea* constitutes the basis of the logic and abstract thought; together with the judgment and argumentation, the complex. The judgment brings together some concepts under certain predication rules (the *sentence* is the *subject predication*), and the argumentation, judgments.

5 The term 'in general' is due to some signs (verbs) where **the word is the significant thing**. In other words, where *decir* (to say) is *hacer* (to do), like to judge, to promise in contrast with *comer* (to eat), *correr* (to run), etc.

6 The **intent** separates symbol or sign or symptom or sign of evidence, not intentional. The same fact can be used as a sign or signal. Let us take for instance the Vatican smoke and that one in a forest, they both 'indicate' that there is fire, but the former is deliberate, intentional (it **means** that there are or not a new Pope, white or black, respectively); in the latter case, it **indicates** that there is fire.

7 We exclude Saussure and Frege, because we consider them to be non functional. We refer to Ogden and Richards, whom are generally accepted in our environment.

d) The thought is defined as *rational speech*, and speech is the language as act. It is an update of the potential or human ability of thinking and talking, which implies to learn the reality and not only to extern it (communicative function).
e) The knowledge is the apprehension of Reality, and this, which includes the principle of identity and space-temporal coordinates, may be defined as *the set of all referents*.
f) Such apprehension of Reality is performed through the language. It relies under the eidetic-conceptual element, in abstraction and generalization of referents (concrete or abstract, folkloric, geometric, cultural or literary), its reduction to the essential characteristics (different to the accidental ones).
g) Finally, but no by coincidence, if we follow the Aristotelian logic and categorical grammar, we will find the overlap between grammatical and logical categories, particularly between *ser*(to be)-*sustantivo*(noun)*; accidente*(accident)-*adjetivo*(adjective)*; acto*(act)-*verbo*(verb).

4 Conclusions

On the basis of the above considerations presented (the thought-language process). This process will depend on the level of system dominion, i.e., to think and link related concepts implies a lexical database and the management of rules in order to link them (grammar). We do not think with images or icons, but with concepts, ideas.

For a simple test, we suggest the reader to think (not to imagine) in a certain reality (abstract or concrete, physical or mental) without appeal to the language (language of any kind) that is no using words.

Research issues on K-means Algorithm: An Experimental Trial Using Matlab

Joaquín Pérez Ortega[1] , Ma. Del Rocío Boone Rojas,[1,2], María J. Somodevilla García[2]

1 Centro Nacional de Investigación y Desarrollo Tecnológico, Cuernavaca Mor. Mex.
2 Benemérita Universidad Autónoma Puebla, Fac. Cs. de la Computación, México.
jperez@cenidet.edu.mx, {rboone,mariasg}@cs.buap.mx

Abstract. Clustering problems arise in many different applications: machine learning data mining and knowledge discovery, data compression and vector quantization, pattern recognition and pattern classification.
It is considered that the k-means algorithm is the best-known squared error-based clustering algorithm, is very simple and can be easily implemented in solving many practical problems.
This paper presents the results of an analysis of the representative works related to the reseach lines of k-means algorithm devoted to overcome its shortcomings. To establish a framework for a proposed improvement to the standard k-means algorithm, the results obtanained from experiments of the k-means in the Matlab package and databases of the UCI repository are presented.

Keywords: k-means, clustering, k-means-Matlab.

1 Introduction.

The problem of object clustering according to its attributes has been widely studied due to its application in areas such as machine learning [4], data mining and knowledge discovery [3, 11], pattern recognition and pattern classification [2]. The aim of clustering is to partition a set of objects which have associated multi-dimensional attribute vectors into homogeneous groups such that the patterns within each group are similar. Several unsupervised learning algoritms have been proposed which partition the set of objects into a given number of groups according to an optimization criterion. One of the most popular and widely studied clustering methods is K-means [10].

This work is part of a project devoted to proposals for improvement to k-means algorithm and its application to health care in México. In the works of [29] and [30] proposes an improvement to k-means algorithm using a new convergence condition and application in cancer incidence by municipalities in México is proposed. Currently, the project focuses on making a proposal for improvement of the algorithm based on the optimization of the classification step. In doing so, we have done an update and analysis of the state of the art in theoretical research of the algorithm. For

purposes of establishing a framework to propose an improvement, a serie of experimental tests on the matlab package for kmeans in databases of the UCI repository has been made.

The works of [41] and [42] highlight the impact and relevance of the k-means algorithm, within the context of Data Mining and clustering techniques. There are over a thousand papers published to date, related to its different forms, applications and contributions of k-means algorithm. Research on the k-means algorithm, has been developed in two major directions, from theory and field of applications. From theory, a set of advantages and shortcomings has been identified, in an attempt of try in to overcome their shortcomings. Also, a serie of works have been developed, that have led different lines of research, on wich this work focuses. Study we have made no attempt to be exhaustive, but rather representative of the reseach. We include in this report, selected works that have been frequently cited and work that was deemed to provide a novel approach. This report is organized as follow: following this introduction, it is included in the Section 2, the approach to the clustering problem and the description of the k-means algorithm. The section 3 provides an analysis and synthesis of the works in the various research lines of k-means. Section 4, describes certain characteristics of k-means in Matlab and in the section 5, we present a reference database of UCI repository and results of test performed to illustrate some aspects of the performance of k-means algorithm on Matlab. Finally, in section 6, we will discuss our preliminary results and present the future work.

2 Clustering Problem and the k-means Algorithm.

According to [15], clusters analysis aims at solving the following very general problem: given a set X of N entities, often described by measurements as points of the real d-dimensional space R^d, find subjets of X wich are homogeneouns and/or well-separated. Homogeneity means that entities in the same cluster must be similar and separation between entities in the same cluster must be similar and separation between entities in different clusters must differ one from the other. These concepts can be made precise in a variety of ways, wich lead to as many clustering problems and even more heuristic or exact algoritms, so clustering is a vast subject.

The work of [42] provides a good survey on the issue of clustering and provides a set of key references..
According to [15], a mathematical programming formulation of the minimum sum-of-squares clustering problema is a follows:

$$\min f(M,Z) = \sum_{i=1}^{M} \sum_{j=1}^{N} z_{ij} \| x_j \quad m_i \|^2$$

subject to

$$\sum_{i=1}^{M} z_{ij} = 1,\ j = 1,2,\ldots,N$$

$$z_{ij} \in \{0,1\} \quad i = 1,2, \ldots, M; \; j = 1,2, \ldots, N$$

$$\text{where} \qquad m_i = \frac{\sum_{j=1}^{N} z_{ij} x_j}{\sum_{j=1}^{N} z_{ij}}, \qquad i = 1,2, \ldots M.$$

The N entities to be clustered are at given points $x_j = (x_{i1}, x_{j2}, \ldots, x_{jd})$ of R^d for j = 1, …., N; M cluster centroids must be located at unknown points $m_i \in R^d$ for i = 1, …, M. The decision variable z_{ij} is equal to 1 if point j is assigned to cluster i, at a squared Euclidean distance $\|x_j - m_i\|^2$ from its centroid. It is well-known that condition $z_{ij} \in \{0,1\}$ may be replaced by $z_{ij} \in [0,1]$, since in the optimal solution, each entity belongs to the cluster with the nearest centroid.

Among the clustering algorithms based on minimizing an objective function or the squared error, perhaps the most widely used and studied is called the k-means algorithm. This algorithm has been discovered by several research across different disciplines, most notably [23], [24] and [12].
And the best-known heuristic for minimum sum-of-squares clustering is MacQuee`s [24]. It proceeds by seleccting a first set of M points as candidate centroid set, then alternately (i) assigning points of X to their closest centroid and (ii) recomputing centroids of clusters so-obtained, until stability is attained.

In a more specific form in the work of [29] four algorithm steps can be identified:

Step 1. Initialization. A set of objects to be partitioned, the number of groups and a centroid for each group are defined.
Step 2. Classification. For each database object its distance to each of the centroids is calculated, the closest centroid is determined, and the object is incorporated to the group related to this centroid.
Step 3. Centroid calculation. For each group generated in the previous step, its centroid is recalculated.
Step 4. Convergence condition. Several convergence conditions have been used from which the most utilized are the following: stopping when reaching a given number of iterations, stopping when there is no exchange of objects among groups, or stopping when the difference among centroids at two consecutive iterations is smaller than a given threshold. If the convergence condition is not satisfied, steps two, three and four of the algorithm are repeated.

3 Analysis and Classification of Papers Reviewed.

This section, identifies the advantages of k-means. The following is the result of the analysis of the work that has been developed in different research lines of k-means, related to the proposals made to try to overcome their shortcomings. The results are organized by category it, has been included in each case, the authors' names, the title and a concise comment on the thrust of the work.

3.1 Algortihm K-means Advantajes.

MacQeen J. [24], the author of one of the initial k-means algorithm and the most frequently cited, states.

The process, which is called "k-means", appears to give partitions which are reasonably efficient in the sense of within-class variance, corroborated to some extend by mathematical analysis and practical experience. Also, the k-means procedure is easily programmed and is computationally economical, so that it is feasible to process very large samples on a digital computer.

Likewise [39], summarizes the benefits of k-means, in the introduction to his work:

K-means algortihm is one of first which a data analyst will use to investigate a new data set because it is algorthmically simple, relatively robust and gives "good enough" answers over a wide variety of data sets.

3.2 Algorithm K-means Shortcomings.

Taking as a framework and as an extension and update, the k-means shortcomings that are identified in [42] the following is the result of the analysis previously cited in a series of tables grouped by category of work that arose as extensions k-means or as possible solutions to one or more of the limitations that have been identified above.

3.2.1 The algorithm's sensitivity to initial conditions: The number of partitions, the initial centroids.

According to [42] there is a universal and efficient method to identify initial patterns and the number k of clusters. In [40] briefly is discussed the sensitivity of the algorithm for the allocation of initial centroids, that in practice the usual method is to test iteratively with a random allocation to find the best allocation in terms of minimizing the total squared distance. However, there have been various investigations aimed at making various proposals related to these limitations:

Authors	Title and Commentary
[45] Zhang, Chen; Xia Shixiong.	"K-means Clustering Algorithm with improved initial Center." It avoids the initial random assignment of centers. Use strategy called "sub-merger"
[2] B. Bahmani Firouzi, T. Niknam, and M. Nayeripour.	"A New Evolutionary Algorithm for Cluster Analysis". It not depend on the initial centers. Algorithm PSO-SA-K combines the algorithms "Particle Swarm Optimization (PSO)," Simulated Annealing "(SA) and K-means.
[39] Barbakh Wesam And Colin Fyfe.	"Local vs global interactions in clustering algorithms: Advances over K-means." It focuses on the algorithm's sensitivity to initial conditions. Incorporate information on the role of overall performance. Define three new algorithms: Weighted k-means (WK), Inverse Weighted K-means (IWK) and Inverse Exponential k-means (IEK).

[19] Kao, Yi-Tung, Zahara, Erwie, Kao. I-Wei.	"A hibridized approach to data clustering". Draft bioinformatics. Hybrid techniques called K-NM-PSO-based K-means, Nelder-Mead Simplex search and optimization of exchange of particles.
[6] Deelers S. And S. Auwatanamongkol.	"Enhancing K-Means Algorithm with Initial Cluster Centers Derived from Data Partitioning along the Data Axis with the Highest Variance." Title explicit.
[34] Redmond, Stephen J., Heneghan, Conor.	A method for initialising the K-means clustering algorithm using kd-trees" A kd-tree used to calculate an estimate of the density of data and to select the number of clusters.
[15] P. Hansen & E. Nagai.	"Analysis of Global k-means, an Incremental Heuristic for Minimum Sum of Squares Clustering". Commentary on work [22].
[31] Pham, D. Dimov, S.S. Nguyen, C.D.	"Selection of K in K-means clustering". It proposes a measure to select the reference number of clusters.
[22] Likas, A., Vlassis, N., Verbeek, J.J.	"The Global K-means Clustering Algorithm." Algorithm that consists of a series of k-means clusterings with varying number of clusters from 1 to k. It argues that it is independent of initial partitions and accelerates the calculations of k-means.
[28] J. Peña, J. Lozano and P. Larrañaga,	"An empirical comparision of four initialization methods for the k-means algorithm." Compare initialization methods for k-means: Random, [12], [20] and [24].
[5] P. Bradley, U.Fayyad.	"Refining initial points for k-means clustering". Use k-means M times for M random subsets of the original data.
[20] L. Kaufman and P. Rouseeuw.	L. Kaufman and P. Rouseeuw. Finding Groups in Data: An Introduction to Cluster analysis: Text. Def. K-means.
[3] G. Ball and D. Hall	"A clustering technique for summarizing multivariate data", (ISODATA). Perform dynamic estimation of K.

3.2.2 The convergence of algorithm to a local optimum rather than a global optimum.

According to [24], the iterative procedure of k-means can not guarantee convergence to a global optimum, but in his work, some research is cited, which are special cases. Currently, there are several developments that analyze and / or proposed solutions to this constraint:

Authors	Título and Commentary
[39] Barbakh Wesam And Colin Fyfe.	"Local vs global interactions in clustering algorithms: Advances over K-means." Addresses the algorithm's sensitivity to initial conditions. Incorporating global information on the performance function. Define three new algorithms: Weighted k-means (WK), Inverse Weighted K-means (IWK) and Inverse Exponential k-means (IEK).
[29] Joaquín .Pérez O, Rodolfo Pazos R, Laura Cruz R.,Gerardo Reyes S. Rosy Basave T. Héctor Fraire H.	"Improvement the Efficiency and Efficacy of the K-means Clustering Algorithm through a New Convergence Condition". Improvement to the k-means algorithm by new convergence conditions. Experimentally analyze the local convergence of k-means.
[44] Z. Zhang, B. Tian D. And Tung A.K.H.	"On the Lower Bound of Local Optimums in K-means Algorithm." Estimate lower limit for local optimum.

[21] K. Krishna and M. Murty	"Genetic K-means algorithm". Hybrid scheme based on Genetic Algorithm - Simulated annealing with new operators to perform global search and rapid convergence.
[24] MacQUEEN J.	"Some Methods for Classification and Analysis of Multivariate Observations." Definition, Analysis and Applications of k-means.

3.2.3 The efficiency of the algorithm.

According to the work of [42] the complexity of the k-means algorithm is O (n, d, k) which involves the sample size, the number of dimensions and the number of partitions. There are several works that have focused on different aspects of the algorithm, in order to reduce computational load.

Authors	Título and Commentary
[4] Moh`d Belal Al-Zoubi, Amjad Hudaib, Ammar Huneiti and Bassam Hammo	"New Efficient Strategy to Accelerate k-Means Clustering Algorithm". Strategy to accelerate k-means algorithm, which avoids many calculations of distance, through a strategy based on an improvement to the partial distance algorithm (PD).
[43] Zalik, Krista Rizman	"An Efficient k`-means Clustering Algorithm." Based on the algorithm Rival, it penalizes competitive Learninig (RPCL). It does not require pre-allocation of the number of clusters. Two-step process. Pre processes and uses the prior information to minimize the cost function.
[26] Cao. D. Nguyen & Cios, Krzysztof J.	"GAKREM: A novel hybrid clustering algorithm." Eliminates the need to specify a priori the number of clusters. Combines genetic algorithms and logarithmic regression Màxima expectation.
[13] G. Frahling & Ch. Sohler.	"A Fast k-means implementation using coresets." Implemented version of Lloyd's k-means [23], using a weighted set of points that approximate the original set.
[35] Taoying Li & Yan Chen	"An improved k-means algorithm for clustering using entropy weighting measures". Improvement of the algorithm by introducing a variable to the function of cost.
[18] Kashima, H. Hu, J.; Ray,B; Singh, M.	"K-means clustering of proportional data using L1 distance". K-means based on distance L1. Proportionate restrictions incorporated in the calculation of centroids.
[36] Tsai, Chieh-Yuan, Chiu, Chuang-Cheng.	"Developing a feature weight self-adjustment mechanism for a K-means clustering algorithm." Improvement of the quality of k-means clustering via FWSA mechanism called "Self-Adjusment Feature Weight." Is modeled as an optimization problem.
[29] Joaquín .Pérez O, Rodolfo Pazos R, Laura Cruz R.,Gerardo Reyes S. Rosy Basave T. Héctor Fraire H.	"Improvement the Efficiency and Efficacy of the K-means Clustering Algorithm through a New Convergence Condition". Improvement to the k-means algorithm by new convergence conditions. Experimentally analyze the local convergence of k-means.
[30] J.Pérez, M.F. Henriques, R. Pazos, L. Cruz, G. Reyes, J. Salinas, A. Mexicano.	"Improvement of the K-means algorithm using a new approach of convergence and its application to databases cancer population." Title explicit.

[33] Pun, W.K.D., Ali, A.S.	"Unique distance measure approach for K-means (UDMA-Km) clustering algorithm." Sets distance measure based on statistical data.
[7] Zejin Ding, Jian Yu, Self-Yang-Qing Zhang.	"A New improved K-Means Algorithm with Penalizaed Term". Define new objective function and minimize it with genetic algorithm.
[17] Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R., Wu, A.Y.:	"An Efficient K-means Clustering Algorithm: Analysis and Implementation," presents an implementation of the version of Lloyd's k-means [23] called "filtering algorithm" based on a kd-tree.

3.2.4 K-means is sensitive to *outliers* and noise.

According to [42] even if an object is quite far away from the cluster centroid, it is still forced into a cluster and, thus, it distorts the cluster shapes. Here are works that focus on shortcoming:

Authors	Título and Commentary
[1] Asgharbeygi, N. Maleki, A.	"Geodesic K-means clustering".Extends k-means by using a geodesic distance metric. Algorithm ensures resistance to outliers.
[9] V. Estivill-Castro and J. Yang	"A fast and robust general purpose clustering algorithm." It eliminates the effect of outliers through a process that considers real points as centroids.
[3] G. Ball and D. Hall	"A clustering technique for summarizing multivariate data".(ISODATA). It performs dynamic estimation of K. Considers the effect of outliers in the process of clustering.

3.2.5 The definition of "means" limits the application only to numerical variables.

Several works have been developed that extend the application of k-means for categorical variables or others:

Authors	Título and Commentary
[38] Song, Wei, Li Cheng Hua, Park, Soon Cheo.	"Genetic Algorithm for text clustering using ontology and evaluating the vality of various semantic simility measures." Improving the k-means algorithm by using a genetic algorithm that finds similarities conceptual. Based on ontology, thesaurus corpus for clustering of text fields.
[14] S. Gupata, K. Rao, &Bhatnagar	"K-means clustering algorithm for categorical attributes". Title explicit.
[16] Z. Huang.	"Extensions to the k-means algorithm for clustering large data sets with categorical values." Title explicit.

4 The Algorithm k-means on Matlab.

Experimental tests were conducted for K-means in the Matlab [25]. The Matlab (Matrix Laboratory) is both, an environment and programming language for numerical calculations with vectors and matrices. It is a product of the company The

Math Works Inc. (Natick, MA). [1]. The K-means algorithm for clustering is in the following MATLAB function:

[IDX, C, SUMD, D] = KMEANS(X, K)

This function partitions the points in the N-by-P data matrix X into K clusters. This partition minimizes the sum, over all clusters, of the within-cluster sums of point-to-cluster-centroid distances. Rows of X correspond to points, columns correspond to variables. KMEANS returns an N-by-1 vector IDX containing the cluster indices of each point. By default, KMEANS uses squared Euclidean distances. The K cluster centroid is located in the K-by-P matrix C. The within-cluster sums point-to-centroid distances in the 1-by-K vector sumD. Distances from each point to every centroid in the N-by-K matrix D. It may include optional parameters to specify distance measure, the method used to choose the initial cluster centroid positions, display information.

5 Test Results of K-means in Matlab.

Tests for k-means in Matlab, used the well known UCI Machine Learning Repository [37]. The UCI Machine Learning Repository [37] is among other things, a collection of databases, which is widely used by the research community of Machine Learning, especially for the empirical algorithms analysis of this discipline.

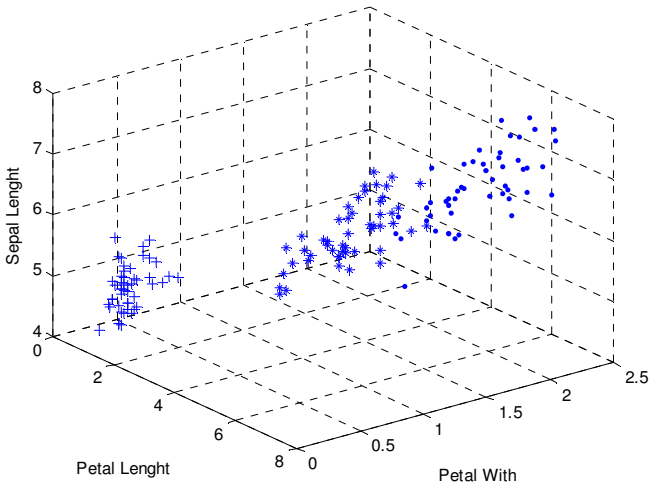


Fig.1 Representation of the Iris Data Set

For the experimental tests carried, the following data sets have been used: Iris, Glass and Wine. This report presents the results for the Iris data set.

The Iris Data Set is a database of types Iris plant, which has No. of instances: 150 (50 in each class), No. of attributes: 4 (Sepal length, Sepal width, Petal Length, Petal width) No. of classes: 3 (Hedges Iris, Iris versicolor, Iris virginica). One class is linearly separable from the other two; the latter are NOT linearly separable from each other. Based on data and classes defined in [37] and [42], fig.1 includes the Iris data, for illustrative purposes only are considered the attributes sepal length, petal length and petal width.

Test I4: >> [u,v,sumd,D]= kmeans(z,3,'display','iter');

iter	phase	num	sum	%inter
1	1	150	426.888	100
2	1	34	134.187	22.6
3	1	13	105.771	8.6
4	1	12	88.8948	8
5	1	6	85.2326	4
6	1	4	84.064	2.6
7	1	3	83.3704	2
8	1	5	82.073	3.3
9	1	3	81.3672	2
10	1	4	80.3157	2.6
11	1	3	79.6817	2
12	1	3	79.1156	2
13	1	1	78.9451	0.6
14	2	1	78.9408	0.6

14 iterations, total sum of distances = 78.9408

The "Test I4" is an example of the test results on Matlab and kmeans for the Iris data set. ***Iter*** column represents the number of iteration, the ***phase*** indicates the algorithm phase, ***num*** provides the number of exchanged points, ***sum***, provides the total sum of distances, ***inter%*** are the percentage of exchanged points in each iteration.

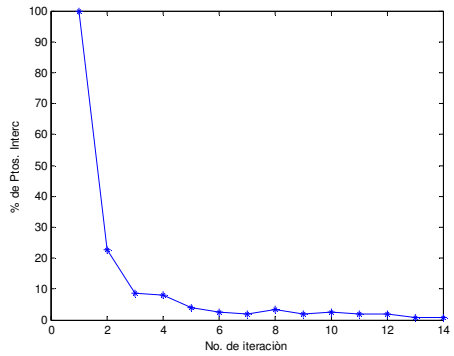


Fig. 2 % of Exchanged Points.

Fig. 2 corresponds to the test I4 and the graphical representation of the exchange behavior at each iteration. Likewise Fig. 3 represents the behavior of the sum of distances for the same test.

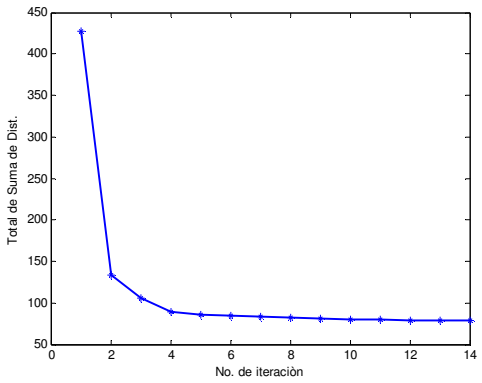


Fig. 3 Total Sum of Distance

No. Prue.	No. Iter.	No. Ptos.Interc.2	% Dif. Ptos. Interc. 1 a 2
1	5	14	90.7
2	10	7	95.4
3	3	5	96.7
4	6	8	94.7
5	14	34	77.4
6	4	5	96.7
7	8	8	94.7
8	8	36	97.6
9	7	73	95.4
10	3	4	97.4
11	7	37	97.6
12	11	9	94.0
13	6	53	64.7
14	5	12	92.0
15	3	2	98.7
16	6	15	90.0
17	4	3	98.0
18	10	9	96.0
19	10	4	97.4
20	6	32	79.0
21	12	16	89.4
22	8	33	78.0
23	7	31	80.0
24	6	14	90.7
25	11	7	95.4

Table 1 Summary of Results for Iris Data Set

5.1 Summary of Results for the Iris data set.

Table 1 provides a summary of 25 experimental tests performed on the IRIS data set, the first column identifies the test number and the second column includes the number of iterations performed by the algorithm.

With regard to exchanges between groups that the algorithm makes in the tests conducted, it was observed that the most significant changes occurred from first to second iteration. For all cases, in the first step all points are located (100%), the third column includes the number of points to be exchanged in the second iteration and the fourth column is the percentage difference in the number of items exchanged between the first and second iteration.

According to the results in Table 1:

It can see that for 150 points in the Iris database, and a set of 25 tests, the k-means algorithm in Matlab:

- ✓ Converge in an average of 7.2 iterations.
- ✓ The average number of points exchanged during the second iteration was 18.84 points
- ✓ The percentage of points located on the second iteration in the corresponding group was 91.0%

6 Conclusions.

The results of the analysis for our sample work, allow us to establish a framework and analyze the theoretical study of the k-means algorithm Also we reseach and distinguish the different lines on which there is still a fertile field for investigation.
As we can see several attempts at overcoming the shortcomings of the k-means algorithm have been done and different approaches in different disciplines have been proposed: Optimization, Probability and Statistics, Neural Networks, Evolutionary Algorithms, among others. The vast majority of contributions have focused on the first three lines of research identified in this study: The sensibility of the algorithm to initial conditions, convergence of the algorithm to a local optimum rather than a global optimum and the efficiency of the algorithm. Notes that challenges still to be resolved in such research and has been relatively little work done on the lines related to the implementation of the algorithm to other variables as well as treatment to outliers and noise.
Acording the tests conducted in Matlab, this laboratory showed that it is actually very conducive to experimental testing, the implementation of k-means, allows to monitor the performance of the algorithm through the information that can be deployed at runtime, such as result of the objective function and the number of points exchanged in each iteration.The results allow us to establish a framework to compare the proposal improvement algorithm with the previous work. As part of this project and to give continuity to previous work [29] [30], also ventures into different applications to k-means, such as in the areas of health care in Mexico and in Web Usage Mining for Log files from the server of the Faculty of Compute Science BUAP, México.

References

1. Asgharbeygi, N. Maleki, A. "Geodesic K-means clustering". Pattern Recognition, 2008, ICPR 2008, 19th International Conference on. Dec. 2008.
2. Bahmani, B., Firouzi, T. Niknam, and M. Nayeripour. "A New Evolutionary Algorithm for Cluster Analysis". Proceedings of world Academy of Science, Engineering and Technology Vol. 36, Dec.2008.
3. Ball, G. and D. Hall, "A clustering technique for summarizing multivariate data", (ISODATA), Behav Sci., vol. 12, pp. 153-155, 1967.
4. Belal Al-Zoubi, Al-Zoubi, Amjad Hudaib, Ammar Huneiti and Bassam Hammo. "New Efficient Strategy to Accelerate k-Means Clustering Algorithm". American Journal of Applied Sciences 5(9) 1247-1250, Science Publications. 2008.
5. Bradley P., U.Fayyad. "Refining initial points for k-means clustering", in Proc. 15 th Int. Conf. Machine Learning, 1998 pp.91-99.
6. Deelers S. And S. Auwatanamongkol. "Enhancing K-Means Algorithm with Initial Cluster Centers Derived from Data Partitioning along the Data Axis with the Highest Variance." Proceedings of world Academy of Science, Engineering and Technology. Vol. 26, Dec. 2007.
7. Ding Zejin, Jian Yu, Yang-Qing Zhang. "A New improved K-Means Algorithm with Penalizaed Term". Granular Computing, 2007, GRC 2007, IEEE International Conference on. Nov. 2007.
8. Duda, R.O., Hart, P.E.: Pattern Classification and Scene Analysis, John Wiley &Sons, New York, NY, 1973.
9. Estivill-Castro V. and J. Yang, "A fast and robust general purpose clustering algorithm." In Proc. 6 th Pacific Rim Int. Conf. Artificial Intelligence (PRICAI`00), R. Mizoguchi and J. Slaney, Eds., Melbourne, Australia, 2000, pp, 208 – 218.
10. Fayyad, U.M., Piatetsky-Shanpiro, G., Smyth P., Uthurusamy, R.: Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996.
11. Fissher, D.: Knowledge Acquisition via Incremental Conceptual Clustering. Machine Learning, Vol.2, No. 2 (1987) 139-172.
12. Forgy E. "Cluster analysis of multivariate data: Efficiency vs. Interpretability of classification", Biometrics, vol. 21, pp.768-780.1965
13. Frahling, G. & Ch. Sohler. "A Fast k-means implementation using coresets.". International Journal of Computational Geometry & Applications. Dec. 2008. Vol. 18 Issue 6. P605-625.
14. Gupata S., K. Rao, and V. Bhatnagar, "K-means clustering algorithm for categorical attributes", in Proc. 1st Int. Conf. Data Werehousing and Knowledge Discovery (DaWak`99). Florence, Italy, 1999, pp. 203 – 208.
15. Hansen, P. & E. Nagai. "Analysis of Global k-means, an Incremental Heuristic for Minimum Sum of Squares Clustering". Journal Classification 22:287-310.
16. Huang, Z., "Extensions to the k-means algorithm for clustering large data sets with categorical values.". Data Mining Knowl. Discov., vol. 2, pp. 283 – 304, 1998.
17. Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R., Wu, A.Y.: An Efficient K-means Clustering Algorithm: Analysis and Implementation.

Pattern Analysis and Machine Intelligence, IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 24, No. 7 (2002) 881-892.
18. Kashima, H. Hu, J.; Ray,B; Singh, M. "K-means clustering of proportional data using L1 distance". Pattern Recognition, 2008, ICPR 2008. International Conference On. Volume Issue, Dec. 2008.
19. Kao, Yi-Tung, Zahara, Erwie, Kao. I-Wei. "A hibridized approach to data clustering". Expert Systems with Applications. Vol. 34 Issue 3. P 1754-1762. Apr.2008.
20. Kaufman L. and P. Rouseeuw. Finding Groups in Data: An Introduction to Cluster analysis: Wiley, 1990.
21. Krishna, K. and M. Murty, "Genetic K-means algorithm". IEEE Trans. Syst., Man, Cybern. B., Cybern., vol. 29, no. 3, pp. 433 – 439, Jun. 1999.
22. Likas, A., Vlassis, N., Verbeek, J.J.: The Global K-means Clustering Algorithm. Pattern Recognition. The Journal of the Pattern Recognition Society. Vol. 36, No. 2 (2003) 451-461
23. Lloyd SP "Least squares quantization in PCM. Unpublished Bell Lab. Tech. Note, portions presented at the Institute of Mathematical statistics Meeting Atlantic City, NJ, sep. 1957.
IEEE Trans. inform, Theory (Special Issue on Quantization), vol. IT-28, pp 129 – 137 Mach 1982.
24. MacQueen, J.: Some Methods for Classification and Analysis of Multivariate Observations. In Proceedings Fifth Berkeley Symposium Mathematics Statistics and Probability. Vol. 1. Berkeley, CA (1967) 281-297.
25. Matworks. http: //www.matworks.com
26. Mehmed, K.: Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons. 2003.
27. Nguyen, Cao. D. & Cios, Krzysztof J. "GAKREM: A novel hybrid clustering algorithm. Information Sciences. Vol. 178 Issue 22, p4205-4227- Nov. 2008.
28. Peña, J. Lozano and P. Larrañaga, "An empirical comparision of four initialization methods for the k-means algorithm. "Pattern Recognit Lett., vol. 20 pp. 1027 – 1040, 1999.
29. Pérez J., Rodolfo Pazos R, Laura Cruz R.,Gerardo Reyes S. Rosy Basave T. Héctor Fraire H. "Improvement the Efficiency and Efficacy of the K-means Clustering Algorithm through a New Convergence Condition". Computational Science and Its Applications – ICCSA 2007 – International Conference Proceedings. Springer Verlag.
30. Pérez, J., M.F. Henriques, R. Pazos, L. Cruz, G. Reyes, J. Salinas, A. Mexicano. Mejora al Algoritmo de *K-means* mediante un Nuevo criterio de convergencia y su aplicación a bases de datos poblacionales de cáncer. 2do Taller Latino Iberoamericano de Investigación de Operaciones, "La IO aplicada a la solución de problemas regionales". México. "In Spanish".
31. Pham, D.T. Dimov, S.S. Nguyen, C.D. "Selection of K in K-means clustering". "Proceedings of the Institution of Mechanical Engineers – Part C – Journal of Mechanical Engineering Science; Vol. 219 Issue 1, p103-109. Jan 2005.
32. Proietti, Guido and Christos Faloutsos. "Analysis of Range Queries on Real Region Datasets Stored Using an R-Tree." IEEE Transactions on Knowledge and Data Engieneering., Vol. 12, No. 5, Sep./Oct. 2000.

33. Pun, W.K.D., Ali, A.S. “Unique distance measure approach for K-means (UDMA-Km) clustering algorithm. TENCON 2007 – 2007 IEEE Region 10 Conference. Oct. 30 2007.
34. Redmond, Stephen J., Heneghan, Conor. “A method for initialising the K-means clustering algorithm using kd-trees”. Pattern Recognition Letters; Vol. 28 Issue 8, Jun. 2007.
35. Taoying Li & Yan Chen “An improved k-means algorithm for clustering using entropy weighting measures”. Intelligent Control and Automation, 2008, WCICA 2008, 7th World Congress on. June 2008.
36. Tsai, Chieh-Yuan, Chiu, Chuang-Cheng. “Developing a feature weight self-adjustment mechanism for a K-means clustering algorithm.” Computational Statistics & Data Analysis. Vol. 52 Issue 10. Jun. 2008.
37. UCI. Asuncion, A. & Newman, D.J. (2007). UCI Machine Learning Repository [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, School of Information and Computer Science.
38. Wei, Song, Li Cheng Hua, Park, Soon Cheo. “Genetic Algorithm for text clustering using ontology and evaluating the vality of various semantic simility measures.” Expert Systems with Applications. Vol. 36, Issue 5, Jul. 2009.
39. Wesan, Barbakh And Colin Fyfe. “Local vs global interactions in clustering algorithms: Advances over K-means.” International Journal of knowledge-based and Intelilligent Engineering Systems 12 (2008).83 – 99.
40. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann Publishers. San Diego, CA (1999)
41. Xindong Wu,, V.Kumar, J. Ross Q., J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng,, B. Liu, P. S. Yu, Z. Zhou, M. Steinbach, D. J. Hand and D. Steinberg. “Top 10 algorthms in data mining”. Knowl Inf Syst (2008). Springer.
42. Xu,_Rui and Donald Wunsch II. Survey of Clustering Algorithms. IEEE Transactions on Neural Networks, Vol., 16, No. 3, May 2005.
43. Zalik, Krista Rizman . “An Efficient k`-means Clustering Algorithm.” Pattern Reconition Letters, Vol. 29, I.9. Pag. 1385-1391. Elsevier 07/2008.
44. Zhang, Z., B. Tian D. And Tung A.K.H. “On the Lower Bound of Local Optimums in K-means Algorithm.” Data Mining 2006, ICDM`06 Sixth International Conference on Data Mining. Dec. 2006.
45. Zhang, Chen; Xia Shixiong. “K-means Clustering Algorithm with improved initial Center.” Knowledge Discovery and Data Mining, 2009. Second International Workshop on. Vol. Issue, 23-25 Jan., 2009.

Image Classification by Texture Segmentation using GAF-SVM

Sergio Manuel Dorantes, Manuel Martín Ortiz, María J. Somodevilla, Jesús Lavalle Martínez, Ivo H. Pineda Torres

Facultad de Ciencias de la Computación, BUAP
sergiomanuel@hotmail.com, {mmartin, mariasg, jlavalle, ipineda}@cs.buap.mx

Abstract. Due to the amount of visual information that currently exists, there is a need to classify it properly. In this paper we present an alternative dual method for image categorization according to their texture content defined as GAF-SVM, this method is based in the use of Gabor Filters (GAF) and Support Vector Machine (SVM). To perform the image classification we rely on filtering techniques for feature extraction mixed with statistical learning techniques to perform the data separation. The experiments were carried out by taking a set of images containing coastal beach scenes and a set of images containing city scenes. A feature vector is obtained from applying a bank of Gabor Filters to the input images; the output feature space is then used as an input to the SVM Classifier. The Support Vector Machine is responsible for learning a model that is capable of separating the sets of input images. Experimental results demonstrate the effectiveness of the proposed dual method by getting the error classification rate to near 9%.

I. INTRODUCTION

The proposal for an alternative method of image classification requires analysis of the methods presented so far concerning the area. Extracting visual information from an image to obtain their most important features is essential for classification tasks; over the years various approaches have been presented regarding this field of study such as: color histograms, region-based classification and gray-level values of raw pixels; though one solution has been to incorporate the texture analysis as main feature descriptor. This is largely due to the fact that most surfaces on images contain some kind of texture. In the recent years, texture analysis has been used for object recognition, image interpretation, image segmentation and classification [1, 2, 6, 8, 9, 10].

In recent papers like [4, 6, 7, 10, 11, 12], texture is been understudy in an isolated manner to evaluate the performance of the proposed algorithms, in some cases it has been used artificial textures, which limits the application area of these methods. Textures are used by the human visual system to separate different objects within scenes as well as surface analysis [11]. Texture can be recognized as an irradiation patterns that are perceptually uniform. Textures can be explained as an efficient

measure to estimate the structural differences of orientation, roughness, smoothness or regularity between different regions of an image [14].

But bring out a formal definition of what a texture really is, it became a subjective topic. As it was mention in [13] the definition of texture is dependent on the purpose for which it is being used and outlines some definitions:

1. The basic pattern and repetition frequency of a texture sample could be perceptually invisible, although quantitatively present. In the deterministic formulation texture is considered as a basic local pattern that is periodically repeated over some area.
2. An image texture may be defined as a local arrangement of image irradiances projected from a surface patch of perceptually homogeneous irradiances.
3. Texture is characterized not only by the grey value at a given pixel, but also by the grey value 'pattern' in a neighborhood surrounding the pixel.

Our proposal is based on the use of natural textures in real world images, for that reason the classification model must deal with more complex images in natural conditions.

The 2-D Gabor filters (2D-GF) have certain properties that make them suitable for textural identification in many ways: 2D-GF have tunable orientation and radial frequency bandwidths, tunable center frequencies, and optimally achieve joint resolution in space and spatial frequency. The demodulated Gabor channel envelopes generally contain only low spatial frequencies which are optimally localized in both domains [16].

Gabor filter based methods have been successfully applied for a variety of machine vision application, such as texture segmentation [10, 11, 12, 15, 16, 18], texture classification [9, 13, 19], iris recognition [21, 22, 23], on-road vehicle detection [17], fingerprint classification [20], and as mentioned in [15] edge detection, object detection, image representation, and recognition of handwritten numerals.

This paper is organized as follows: in section II it is mention the related work we based on to develop this article, in section III it is made a detail description of the proposed method, in section IV it is an explanation of the way the input data is processed as well as the Gabor filter's parameters selection, in section V the details of the SVM classifier parameters, and in section VI the experimental results.

II. RELATED WORK

The classification of images has been studied from various approaches, most of all through the mixing of methods, one for texture extraction and one for the classification process.

In [9] is emphasized the use of Gabor filters as a texture extraction method and classification is performed with maximum likelihood method for the classification of aerial and satellite digital images. In [3] is proposed a method of image classification using as an image representation their color histogram and as method of classification the Support Vector Machine. In [4], is not used an external feature extractor, instead the SVM classifier receives the grey level values of each pixel on the image, trying to

prove that SVM can implement feature extraction methods within its architecture, this method is computationally expensive due to the number of regions that can define an image. Another approach is performed in [5] where a modification of the SVM is used for identification of regions among a group of images. In [6] the SVM is combined with the Discrete Wavelet Frame Transform for the classification of images of the Brodatz album. In [7] is mixed the use of wavelet transform as a feature extractor known as the pyramid-structured wavelet transform and SVM as the classification method. In [8] is proposed a method called Gaussian Mixture Model mixed with Independent Component Analysis (ICA) to perform the image classification, which is called ICA Mixture Model.

The first step to complete the proposed method is to extract the texture features with a bank of Gabor filters applied to each input image, and then take the filter's output to form a training dataset to feed the SVM classifier.

III. PROPOSED METHOD

In order to accomplish the image classification we rely on filter based techniques to perform texture feature extraction mixed with statistical learning theory techniques to achieve the image data separation. Gabor filters were selected to extract texture features from images due to their resemblance to the human visual system [13].

A. Gabor Filters

A number of authors have used a bank of filters to extract local images features [10, 11, 16, 19]. Different authors used different sets of Gabor Filters, from spatial domain to frequency domain.

A 2-D Gabor filter is a linear filter whose impulse response is defined by a harmonic function multiplied by a Gaussian function. In the spatial domain can be defined as follows:

$$\psi(x,y) = \frac{f^2}{\pi\gamma\eta} e^{-\left(\frac{f^2}{\gamma^2}x'^2 + \frac{f^2}{\eta^2}y'^2\right)} \cdot e^{j2\pi fx'} \tag{1}$$

$$x' = x\cos\theta + y\sin\theta$$

$$y' = -x\sin\theta + y\cos\theta$$

Where f is the central frequency of the filter, θ the rotation angle of the Gaussian major axis and the plane wave, γ the sharpness along the major axis, and η the sharpness along the minor axis (perpendicular to the wave). In the given form, the aspect ratio of the Gaussian is $\lambda=\eta/\gamma$. This four parameters (f,θ,γ,η) define the shape of the filter, and by changing them we can detect different textures.

The normalized 2-D Gabor filter function has an analytical form in the frequency domain.

$$\Psi(u,v) = e^{-\frac{\pi^2}{f^2}(\gamma^2(u'-f)^2+\eta^2 v'^2)} \quad (2)$$

$$u' = u\cos\theta + v\sin\theta$$

$$v' = -u\sin\theta + v\cos\theta$$

B. Filter Design vs. Filter Bank

There exist two aspects regarding the implementation of Gabor filters, on one hand the filter bank approach and in the other hand the filter design approach [25]. In the first one, a bank of filters is formed by grouping multiple filters tuned at different frequencies and different orientations. The decision of the parameters setting depends on the type of texture to be analyzed. The difficulty of using the filter bank approach relies on the fact that their parameters are established ad hoc and are not optimal for a specific processing task. One of the goals of this work consists on presenting results that would help to specify such parameters. Furthermore, if the bank handles many frequencies and orientations, resulting in a large bank with a lot of filters within, this translates in a large number of convolutions. The filter design approach focuses on designing one or a few filters for a particular application in an effort to reduce the difficulty provided by the filters bank and also to reduce the dimensionality of the output, as well as the processing cost. The disadvantage of this approach lies in the limitation of the tasks for which it was designed. When working with a single filter it is possible that some of the textures in the images are not identified or detected as the filter has a narrow range capacity to detect local texture features.

A filters bank allows the analysis of an image in a single pass way at several frequencies and in several orientations at once. According to the characteristics of our model, the use of a filters bank is the solution choice of deployment, although it could mean an increase, in the computational processing, this is not significant. The design of a Gabor filter bank consists, in general, in the selection, for each filter, of the proper values of the following parameters: *frequency, orientation,* γ and η, the last two parameters known as the smoothing parameters [26].

In this research it is defined a bank with up to 3 orientations and up to 2 frequencies, resulting in a bank with maximum 6 output filters, allowing us to accurately detect a texture among a large set of images. This decision was made based on the studies presented in [26], where is compared with various parameter selection approaches, and summarizes some parameter values adopted in literature.

Using many different orientations and scales (frequencies) ensures invariance; objects and some textures can be recognized al various different orientations, scales and translations [27].

C. Support Vector Machines

Support Vector Machines (SVM) were introduced by Vapnik as a powerful learning tool based on statistical learning theory, a Support Vector Machine is a binary

classifier that makes its decision by constructing a linear decision boundary or hyperplane that optimally separates data points of the two classes in feature hyper space and also makes the margin maximized [20].

SVM starts from the goal of separating the data with a hyperplane, and extend this to non-linear decision boundaries using the kernel trick [29]. A hyperplane can be defined as:

$$w^T x + b = 0 \quad \textbf{(3)}$$

Where x represents a point (a vector), w represent the weight (also a vector). We want to choose w and b to maximize the margin, or distance between the parallel hyperplanes that are as far as possible while still separating the data. The hyperplane must separate data such as:

$$\begin{aligned} w^T x_k + b > 0 \text{ for all } x_k \text{ of a class } y \\ w^T x_j + b < 0 \text{ for all } x_j \text{ of another class} \end{aligned} \quad \textbf{(4)}$$

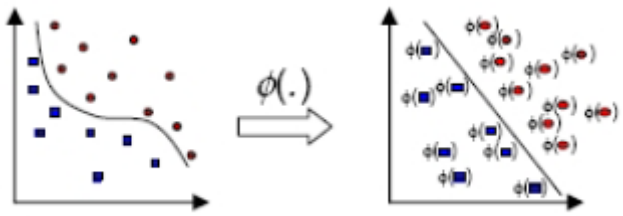
If data are separable in this way, there will probably be more than one way to do it. Among all the possible existing hyperplanes, SVM selects the one in which the distance between the hyperplane and the closest data is the widest possible [29].

When working with a dataset that is not linearly separable, it is necessary to turn to the use of a kernel function. The kernel function allows the SVM to form non-linear boundaries [29]. Data representation through kernel function offers an alternative solution to the nonlinearity problem, projecting the information to a higher dimension feature space [28]. This is accomplished by changing the representation of the function; this is similar to mapping the input space X to a new space H, called feature space, in the form:

$$\phi : X \subset \mathrm{R}^d \rightarrow H \quad \textbf{(5)}$$

Now instead of considering the input vectors $\{x_1, \ldots, x_n\}$ it is considered the transformed vectors $\{\varphi(x_1), \ldots, \varphi(x_n)\}$ as shown in figure 1. By doing this substitution, it is obtained a SVM raised in a new space (this is called the 'kernel trick'), it is important to mention that in practice the implementation of this nonlinear technique consumes the same amount of computational resources of its linear equivalent.

Fig. 1. Using the Kernel to transform (map) the input data space.



The general problem that SVM want to resolve is to search, for a given learning task, with a finite amount of data, an appropriate function that helps to carry out a

good generalization, which results from a proper relationship between the accuracy achieved with a particular training set and the ability of the model [30].

The use of the 'Radial Basis Function' (RBF) kernel is based on the fact that this kernel is basically suited best to deal with data that have a class-conditional probability distribution function approaching the Gaussian distribution, like the texture present on the input images. It maps such data into a different space where the data becomes linearly separable. The kernel function is defined as follows:

$$k(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2} \right) \quad (6)$$

A disadvantage concerning this kernel is that is difficult to design, in the sense that it is difficult to obtain an optimum value for its parameter σ (sigma) and choose the corresponding C that works best for a given problem. The fact that certain combinations of σ and C make the SVM highly sensitive to training data also contributes to the error rate of the RBF-based SVM.

One of the advantages of the RBF kernel is that given the kernel, the weights, the number of support vector and the support vectors itself are automatically obtained as part of the training procedure, i.e. they don't need to be specified by the training mechanism.

IV. SETTING THE EXPERIMENTS

As part of the experiments it was decided to work with two sets of images, one set consisting of coastal beach scenes, and the other set consisting of city scenes images. The processing of input images is done in order to reduce computational complexity.

The first set is conformed by 128 images of beach scenes content, the second set is conformed by 128 images of city scenes content, a total of 256 images.

The input images after processed end up being 8-bit per pixel grayscale images of dimension *128*x*128*, working with just one channel reduces de number of convolutions. The output of each filter is obtained by the convolution of the input image with a Gabor filter. The process is shown below:

$$G(x, y) = I(x, y) \otimes \psi(x, y) \quad (7)$$

where

$G(x, y)$ is the output of the filter

$I(x, y)$ is the original image

$\psi(x, y)$ is the Gabor filter

This computation can theoretically be done in the spatial domain however the Gabor filter is usually narrow. The filter is usually much larger in the frequency domain and thus less affected by aliasing effects due to sampling. It is thus more convenient to do all the computation process in the frequency domain. The convolution is then reduced to a simple and efficient point-wise multiplication of the Fourier transforms [11].

The family of Gabor filters selected to set up the filter bank for the experiments in the frequency domain are:

$$\Psi(u,v) = e^{-\frac{\pi^2}{f^2}(\gamma^2(u'-f)^2+\eta^2 v'^2)} \quad \textbf{(8)}$$
$$u' = u\cos\theta + v\sin\theta$$
$$v' = -u\sin\theta + v\cos\theta$$

A filters bank is constructed by changing values on the four parameters mentioned before, (f,θ, γ, η). For our experiments we only change of the parameters f and θ, which are the central frequency of the filter and the rotation angle of the filter, the other two parameters (γ and η) are set to be constant.

For the experiments it has been set up the bank with 2 frequencies and 3 orientations, to obtain a total of 6 filter outputs for each image on the respective set.

The selected parameters for the filter banks are: central frequency $f = 0.1725$, which outputs the two frequencies 0.1725, 0.1220. Three orientations θ = 0°, 60° and 120°, sharpness along the mayor axis $\gamma=0.5$, sharpness along the minor axis $\eta=0.5$.

In order to apply the filters bank to the sets of images, each image needs to be transform into frequency domain via the Fourier transform. The 2-D Fourier transform used for images can be defined as:

$$I(u,v) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} i(x,y)e^{-i2\pi(ux+vy)}dxdy \quad \textbf{(9)}$$

where

$i(x,y)$ is the original input image

The convolution of the input image with the Gabor filter is performed. In domain frequency the convolution is represented in a point-to-point multiplication of the transformed image with the Gabor filter.

$$g(x,y) = i(x,y)\otimes\psi(x,y) \text{ - Spatial Domain} \quad \textbf{(10)}$$
$$G(x,y) = I(x,y)\cdot\Psi(x,y) \text{ - Frequency Domain}$$

Once the filter output is obtained, *G(x,y)* needs to be transformed back to its spatial representation using the Inverse Fourier Transform in 2-D.

$$i(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I(u,v)e^{i2\pi(ux+vy)}dudv \quad \textbf{(11)}$$

where

$I(u,v)$ is the image in frequency domain

After the transformation, normalization is applied to the output image in order to avoid effects by illumination.

At the end of normalization, we have a certain number of square matrices per filtered image; each matrix dimension is *128*x*128*. The number of square matrices depends on the number of orientations and frequencies concerning the filters bank, in our experiments the filter bank consist of 2 frequencies and 3 orientations, so the number of output matrices is 6.

When convolution is performed some results are not useful especially if the image does not contain textures that respond meaningfully to the filter selected parameters. To reduce the problem all the outputs obtained by convolution are summed up to remove the results that are not relevant and to enhance those that helps to detect texture regions; this also helps to reduce dimensionality of the feature space by having one square matrix as an output, same size of the input image.

At this point we have one matrix per input image, reducing the dimensionality of input data. Each matrix is used to build up a *feature matrix*, which is going to serve as an input of the SVM classifier.

To complete the convolution of the input image with each one of the Gabor filters we take only the real part of the output filtered image. As mentioned in [31], by this way we can keep most the texture response information ignoring phase information.

$$\mathrm{Re}(G(x, y)) \tag{12}$$

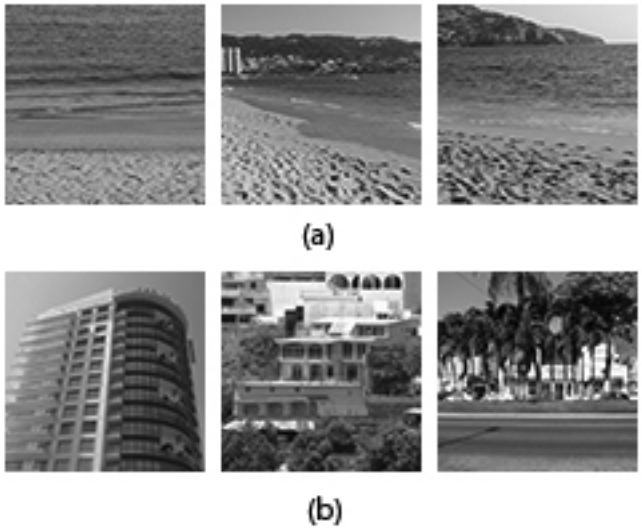
Then we modified each output matrix of dimension *128*x*128* to construct the *feature matrix*. We take each matrix and transform it in a *1*x*16384* vector, each vector is then piled up with the next transformed matrix to form the *feature matrix*.

Finally we have a *feature matrix* of dimension *256*x*16384* which serves as input to the classifier.

V. SVM CLASSIFIER

The goal of experimentation is to obtain a training model through SVM, which can be capable of separate a set of input images. Once we have the *feature matrix* with processed and filtered images we proceed with the SVM classification procedure. According to the nature of the classification process we need to define a training dataset, so the classifier could learn a model, and a test dataset, that let us test the learned model. The training dataset is conformed by 75% of the input dataset, and the test dataset by the remaining 25%. The selection criteria to build up the training dataset and the test dataset are done randomly. In fig. 2(a) is shown an example of coastal beach scene images, and in fig 2 (b) an example of city scenes images, which the classifier will try to separate.

Fig. 2. Example of images used in the experiments. (a) Beach scene images, (b) City images.



The experiments were performed using SPIDER [32], a MATLAB implementation of SVM, a complete object oriented environment for machine learning. Being SVM a binary classifier it is necessary to label the datasets for the classification experiments, the beach scenes images are labeled as 1, and the city scene images are labeled as -1.

In table I there is a list of kernels available in SPIDER, its formula and its parameters.

Table 1. List of Kernel opeators available in SPIDER

Kernel	Parameter	Formula
Linear		$k(x,y) = x \cdot y$
Poly	d, degree	$k(x,y) = (x \cdot y + 1)^d$
Radial Basis Function (RBF)	sigma	$k(x,y) = e^{\frac{-\|x-y\|^2}{2\cdot\sigma^2}}$
Gaussian	sigma	$k(x,y) = \frac{1}{2\pi^{N/2} \cdot \sqrt{\sigma}}$

VI. EXPERIMENTAL RESULTS

It is used the “RBF” kernel to execute the experiments with different sigma values. Another parameter used by SPIDER is the ‘soft margin parameter’, *C*, which penalizes the training errors. This value is set to 1000 in all the experiments.

Iteratively, the sigma values were changed until a significant error rate is obtained. The test results for the learned algorithm are presented in table II.

As it can be seen in table II the sigma value which represents the lower percentage error is $\sigma = 35$, with an error rate of 9.37%.

Table 2. Experimental results for different sigma values

RBF	error	RBF	error
σ=21	**0.1406**	σ=29	**0.1094**
σ=22	**0.1094**	σ=30	**0.1094**
σ=23	**0.1094**	σ=31	**0.1094**
σ=24	**0.1094**	σ=32	**0.1094**
σ=25	**0.1094**	σ=33	**0.1094**
σ=26	**0.1094**	σ=34	**0.1094**
σ=27	**0.1094**	σ=35	**0.09375**
σ=28	**0.1094**	σ=36	**0.09375**

VII. CONCLUSIONS

Extracting texture features by a Gabor filter bank and classify the filter outputs via the Support Vector Machines offers an excellent accuracy rate, 90.63% of the input images are correctly classified according to their class, belonging to beach scenes class or city scenes class.

The article proves the efficiency of using a dual model, first to extract de texture features and then classify them with SVM.

REFERENCES

[1] T. Randen and J.H. Husoy, "Filtering for texture classification: a comparative study", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol 21, Issue 4,, pp. 291 – 310, Apr 1999.

[2] F. Lumbreras Ruiz, "Segmentation, classification and modelization of textures by means of multiresolution decomposition techniques", Ph.D. dissertation, Dept. Informática and Computer Vision Center, Universitat Autònoma de Barcelona, Barcelona, España, 2001.

[3] O. Chapelle, P. Haffner, and V.N. Vapnik, "Support vector machines for histogram-based image classification", *IEEE Trans. On Neural Networks*, Vol. 10, Issue 5, pp. 1055 – 1064, Sep 1999.

[4] Kwang In Kim, Keechul Jung, Se Hyun Park, and Hang Joon Kim, "Support vector machines for texture classification", *IEEE Trans. On Pattern Analysis and Machine Intelligence*, Vol. 24, Issue 11, pp. 1542 – 1550, Nov 2002.

[5] I. Gondra and D.R. Heisterkamp, "Learning in region-based image retrieval with generalized support vector machines", *In Proc. of the Computer Vision and Pattern Recognition*, pp. 149 – 154, 2004.

[6] Shutao Li, J.T. Kwok, Hailong Zhu, and Yaonan Wang, "Texture classification using the support vector machines", *Pattern Recognition*, Vol. 36, No. 12, pp. 2883 – 2893, 2003.

[7] Bing-Yu Sun and De-Shuang Huang, "Texture classification based on support vector machine and wavelet transform", *In Proc. of the Fifth World Congress on Intelligent Control and Automation*, WCICA 2004. Vol. 2, pp. 1862 – 1864, June 15–19, 2004.

[8] V.P. Subramanyam Rallabandi and S.K. Sett, “Unsupervised texture classification and segmentation”, *Proceedings Of World Academy of Science, Engineering and Technology*, Vol. 5, April 2005.
[9] J.A. Recio, L.A. Ruiz and A.Fernández-Sarriá, “Use of Gabor filters for texture classification of digital images”, Física de la Tierra, Vol. 17, pp. 47 – 59, 2005.
[10] M.R. Turner, “Texture discrimination by Gabor functions”, Biol. Cybern., Vol. 55, Num. 2–3, pp. 71 – 82, 1986.
[11] V. Levesque, “Texture segmentation using Gabor filters”, Center for Intelligent Machines Journal, 2000
[12] P. Guha and R. Banerjee, “Segmentation and classification of multi-textured images”, 2000, Available: http://www.cse.iitk.ac.in/~amit/courses/768/00/rajrup/, last visited: April 20, 2009.
[13] V.S. Vyas and P. Rege, “Automated texture analysis with Gabor filters”, GVIP Journal, Vol. 6, Issue 1, pp. 35 – 41, July 2006.
[14] K.M. Rajpoot and N.M. Rajpoot, “Wavelets and Support Vector Machines for Texture Classification”, *In proceedings of the 8th International Multitopic Conference*, INMIC 2004, 24-26 Dec., pp. 328 – 333, 2004.
[15] D.M. Tsai, “Optimal Gabor filter design for texture segmentation”, Technical Report, Machine Vision Lab, Dept. of Ind. Eng. and Mgmt., Yuan-Ze University, Chung-Li, Taiwan, 2000.
[16] A.C. Bovik, M. Clark and W.S.Geisler, “Multichannel Texture Analysis Using Localized Spatial Filters”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 12, Num. 1, pp. 55 – 73, 1990.
[17] Zehang Sun, G. Bebis and R. Miller, “ On-road vehicle detection using Gabor filters and support vector machines”, *14th International Conference on Digital Signal Processing*, DSP 2002, Vol. 2, pp. 1019 – 1022, 2002.
[18] K. Hammouda and E. Jernigan, “Texture segmentation using Gabor filters”, tech. rep., Biotechnology and health engineering centre, University of Waterloo, Dec. 2000.
[19] S.E. Grigorescu, N. Petkov and P. Kruizinga, “Comparison of texture features based on Gabor filters”, *IEEE Trans. On Image Processing*, Vol. 11, Num. 10, pp. 1160 – 1167, 2002.
[20] D. Batra, G. Singhal and S. Chaudhury, “Gabor filter based fingerprint classification using support vector machines”, *Proceedings of the IEEE First India Annual Conference, 2004, INDICON 2004*, pp. 256 – 261, 20-22 Dec. 2004.
[21] Q.A. Salih and V. Dhandapani, “IRIS Recognition based on multi-channel feature extraction using gabor filters”, *Proceedings of the 2nd IASTED international conference on Advances in computer science and technology*, ACST’06, pp. 168 – 173, 2006.
[22] L. Ma, Y. Wang and T. Tan, “Iris recognition based on multichannel Gabor filtering”, 5th Asian Conf. Computer Vision, Vol. 1, 2002.
[23] D. Carr, “Iris recognition: Gabor filtering”, Connexions. Dec. 18, 2004, Available: http://cnx.org/content/m12493/1.4/, last visited April 20, 2009.
[24] K. Kämäräinen, “Feature extraction using Gabor filters”, Ph. D. dissertation, Lappeenranta University of Technology, Finland, Nov. 2003.
[25] T.P. Weldon, W.E. Higgins and D.F. Dunn, “Gabor filter design for multiple texture segmentation”, *Optical Engineering*, Vol. 35, pp. 2852 – 2863, 1996.
[26] F. Bianconi and A. Fernández, “Evaluation of the effects of Gabor filter parameters on texture classification”, *Pattern Recognition*, Vol. 40, Num. 12, pp. 3325 – 3335, 2007.
[27] J. Ilonen, J.K. Kämäräinen and J.K. Kälviäinen, “Efficient computation of Gabor features”, Research Report 100, Lappeenranta University of Technology, Dept. of Information Technology, 2005.
[28] J.A. Reséndiz, “Las máquinas de vectores de soporte para identificación en línea”, Masters dissertation, Departamento de control automático, Centro de investigación y estudios avanzados, I.P.N., 2006.
[29] J.P. Lewis, “A short SVM (support vector machine) tutorial”, CGIT Lab / IMSC, University Southern California, 2004.
[30] L. González, “Modelos de clasificación basados en máquinas de vectores de soporte”, *Asoc. científica europea de econ. aplicada*. Anales de economía aplicada, 2003.
[31] D. Dunn, W.E. Higgins and J. Wakeley, “Texture segmentation using 2-D Gabor elementary functions”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 16, Num. 2, pp. 130 – 149, Feb 1994.
[32] SPIDER, A complete object oriented environment for machine learning in MATLAB. Available: http://www.kyb.mpg.de/bs/people/spider/, last visited May 15, 2009.