Advances on Semantic Web and New Technologies

July, 2010

Editors:

Dra. María Josefa Somodevilla García

Dra. Darnes Vilariño Ayala

Dr. David Eduardo Pinto Avedaño

The Workshop on Semantic Web and New Technologies was held by third time at the Faculty of Computer Science of Benemérita Universidad Autónoma de Puebla, Mexico in July 2010.

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.

Davide Buscaldi and Gerardo Sierra were the invited speakers in this Third Workshop Semantic Web.

Davide Buscaldi is currently completing his Ph.D. in pattern recognition and artificial intelligence at the UPV - Universidad Politécnica de Valencia (Spain), with a thesis titled "Toponym Disambiguation in NLP Applications". His research interests are mainly focused on question answering, word sense disambiguation and geographical information retrieval. He obtained his DEA (Diploma de Estudios Avanzados) in 2008 with a dissertation on the "integration of resources for QA and GIR". He is the author of over 40 papers in different international conferences, workshops and journals. He has been awarded a FPI grant by the Valencian local government which allowed him to participate in the "LiveMemories" project during a stage at the FBK-IRST research institute in Trento, Italy, under the direction of Bernardo Magnini. He has been the UPV responsible of the organization of the QAST (Question Answering on Speech Transcript) track in CLEF 2009. Currently, he is member of the Natural Language Engineering (NLE) Lab of the Universidad Politécnica de Valencia.

Gerardo Sierra is a Ph.D. in Computational Linguistics at UMIST, England. He is the coordinator of the Linguistic Engineering Group at UNAM. He has promoted this area in teaching level such as research and development, in areas such as computational lexicography, terminotics, retrieval and information extraction, text mining and corpus linguistics. Currently, he is researcher level A, National Researcher II, CONACYT project evaluator, member of several scientific committees. He has taught courses at UNAM, for the Faculties of Engineering and Philosophy and Letters, such as Posgrade in Linguistic, Biotechnology and Computer Science.

Content

Ambiguous Place Names on the Web*

Davide Buscaldi

Natural Language Engineering Lab., ELiRF Research Group,
Dpto. de Sistemas Informáticos y Computación (DSIC),
Universidad Politécnica de Valencia, Spain,
dbuscaldi@dsic.upv.es

Abstract. Geographical information is achieving an increasing importance in the World Wide Web. Everyday, the number of users looking for geographically constrained information is growing. Map-based services, such as Google or Yahoo Maps provide users with a graphical interface, visualizing results on maps. However, most of the geographical information contained in web documents is represented by means of toponyms, which in many cases are ambiguous. Therefore, it is important to properly disambiguate toponyms in order to improve the accuracy of web searches. The advent of the semantic web will allow to overcame this issue by labelling documents with geographical IDs. In this paper we discuss the problems of using toponyms in web documents instead of identifying places using tools such as Geonames RDF, focusing on the errors that affect a prototype geographical web search engine, Geooreka!, currently under development.

1 Introduction

The interest of users for geographically constrained information in the Web has increased over the past years, boosted by the availability of services such as Google Maps[1]. Sanderson and Kohler [1] showed that 18.6% of the queries submitted to the Excite search engine contained at least a geographic term, while Gan et al. [2] estimated that 12.94% of queries submitted to the AOL search engine expressed a geographically constrained information need. Most of the geographical information contained in the Web and unstructured text is composed by *toponyms*, or place names. There are two main problems that derive from using toponyms to represent geographical information. The first one is the polysemy of toponyms, or toponym ambiguity: a toponym may be used to represent more than one place, such as "Puebla" which may be used to indicate the city at $19^o3'$N, $98^o12'$W, the state in which it is contained, a suburb of Mexicali in the state of Baja California, or three more small towns in Mexico. The second problem is that the mere inclusion of a toponym in a document does not always mean that the document is geographically relevant with respect to the region or

* We would like to thank the TIN2009-13391-C04-03 research project for partially supporting this work.

[1] http://maps.google.com

area represented by the toponym. In the first case, the solution is constituted by the *Toponym Disambiguation* (TD) task, also named toponym grounding or resolution; in the second case, the solution is to carry out *Geographic Scope Resolution*, which is also affected by the problem of toponym ambiguity [3].

The Geonames ontology[2] provide users with RDF description of more than 6 million places. The use of this ontology would allow to include geospatial semantic information in the Web, eliminating the need of toponym disambiguation. Unfortunately, as noted by [4], in the Web "references to geographical locations remain unstructured and typically implicit in nature", determining a "lack of explicit spatial knowledge within the Web" which "makes it difficult to service user needs for location-specific information". In this paper, with the help of the Geooreka![3] system [5], a prototype web search engine developed at the Universidad Politécnica of Valencia in Spain, we will the problems that users interested in geographically constrained information may found because of the ambiguity of toponyms in the web.

2 Geooreka!: a Geographical Web Search Engine

Geooreka! is a search engine developed on the basis of our experiences at GeoCLEF[4] [6,7], which suggested us that the use of term-based queries could not be the optimal method to express a geographically constrained information need. For instance, it is common for users to employ vernacular names that have vague spatial extent and which do not correspond to the official administrative place name terminology. Another issue is the use of vague geographical constraints that are difficult to automatically translate from the natural language to a precise query. For instance, the query "Cultivos de tabaco al este de Puebla" ("Tobacco plantations East of Puebla") presents a double problem because of the ambiguity of the place name and the fact that the geographical constraint "East of" is vague (for instance, it does not specify if the search should be constrained within Mexico or extend to other countries).

These issues are addressed in Geooreka! by allowing the user to specify his geographical information needs using a map-based interface. The user writes a natural language query in order to represent the query theme (e.g., "Cultivos de tabaco") and selects a rectangular map in a box (Figure 1), representing the query geographical footprint. All toponyms in the box are retrieved using a PostGIS database, and then the Web is queried in order to check the maximum Mutual Information (MI) between the thematic part of the query and all the places retrieved. The complete architecture of the system can be observed in Figure 2. Web counts and MI are used in order to determine which combinations theme-toponym are most relevant with respect to the information need expressed by the user (*Selection of Relevant Queries*). In order to speed-up the process,

[2] http://www.geonames.org/ontology/
[3] http://www.geooreka.eu
[4] http://ir.shef.ac.uk/geoclef/

Fig. 1. Main page of Geooreka!

web counts are calculated using the static Google 1T Web database[5], indexed using the jWeb1T interface [8], whereas Yahoo! Search is used to retrieve the results of the queries composed by the combination of a theme and a toponym.

2.1 Model of Theme-Place Relevance

The key issue in the selection of the relevant queries is to obtain a relevance model that is able to select pairs theme-toponym that are most promising to satisfy the user's information need. On the basis of the theory of probability, we assume that the two component parts of a query, theme T and a place G, are independent if their conditional probabilities are independent, i.e., $p(T|G) = p(T)$ and $p(G|T) = p(G)$, or, equivalently, their joint probability is the product of their probabilities:

$$\hat{p}(T \cap G) = p(G)p(T) \tag{1}$$

If probabilities are calculated using page counts, that is, as the number of pages in which the term (or phrase) representing the theme or toponym appears, divided by $F_{max} = 2,147,436,244$ which is the maximum term frequency contained in the Google Web 1T database, then $\hat{p}(T \cap G)$ is the *expected* probability of co-occurrence of T and G in the same web page. It is clear that this represents a rough estimation of the fact that T occurred in G, since the mere inclusion of G in a page where T is mentioned does not guarantee the semantic relation between G and T.

Considering this model for the independence of theme and place, we can measure the divergence of the expected probability $\hat{p}(T \cap G)$ from the observed probability $p(T \cap G)$: the more the divergence, the more informative is the result

[5] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13

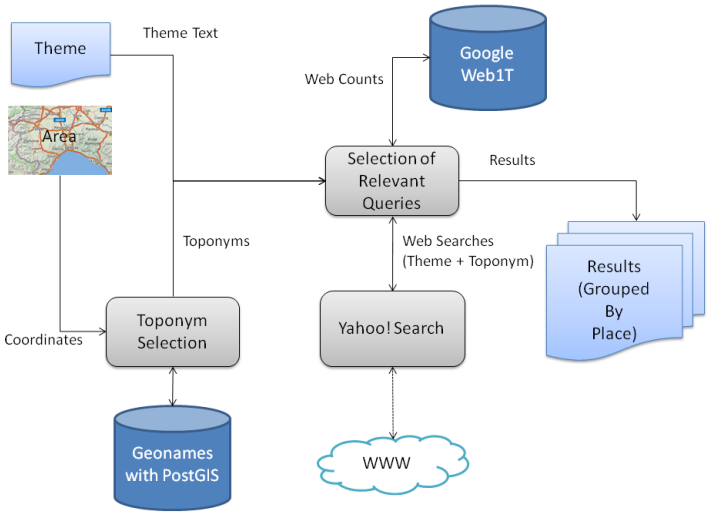

Fig. 2. Architecture of Geooreka!

of the query. The Kullback-Leibler measure [9] is commonly used in order to determine the divergence of two probability distributions.

$$D_{KL}(p(T \cap G)||\hat{p}(T \cap G)) = p(T \cap G) \log \frac{p(T \cap G)}{p(T)p(G)} \quad (2)$$

This formula is exactly one of the formulations of the *Mutual Information* (MI) of T and G, usually denoted as $(I(T;G))$.

3 Evaluation

Geooreka! has been evaluated over the GeoCLEF 2005 test set, in order to compare the results that could be obtained by specifying the geographic footprint by means of keywords and those that could be obtained using a map-based interface to define the geographic footprint of the query. With this setup, topic title only was used as input for the Geooreka! thematic part, while the area corresponding to the geographic scope of the topic was manually selected. Probabilities were calculated using the number of occurrences in the GeoCLEF collection. Occurrences for toponyms were calculated by taking into account only the *geo* index. The results were calculated over the 25 topics of GeoCLEF-2005, minus the queries in which the geographic footprint was composed of disjoint areas (for instance, "Europe" and "USA" or "California" and "Australia"), which could not be processed by Geooreka!. Mean Reciprocal Rank (MRR) was used as a measure of accuracy. The GIR system GeoWorSE, where queries are specified by text, was used as a baseline [10]. Table 1 displays the obtained results.

Table 1. MRR obtained with Geooreka!, using GeoCLEF or the WWW as target collection, compared to the MRR obtained using the GeoWorSE system, Topic Only runs.

topic	GeoWorSE	Geooreka! (GeoCLEF collection)	Geooreka! (Web)
GC-002	0.250	1.000	0.083
GC-003	0.013	1.000	1.000
GC-005	1.000	1.000	0.000
GC-006	0.143	0.000	0.500
GC-007	1.000	1.000	0.125
GC-008	0.143	1.000	0.000
GC-009	1.000	1.000	0.067
GC-010	1.000	0.333	0.250
GC-012	0.500	1.000	0.000
GC-013	1.000	0.000	0.000
GC-014	1.000	0.500	0.091
GC-015	1.000	1.000	1.000
GC-016	0.000	0.000	1.000
GC-017	1.000	1.000	0.143
GC-018	1.000	0.333	0.500
GC-019	0.200	1.000	0.045
GC-020	0.500	1.000	0.090
GC-021	1.000	1.000	0.000
GC-022	0.333	1.000	0.076
GC-023	0.019	0.200	0.125
GC-024	0.250	1.000	1.000
GC-025	0.500	0.000	0.000
average	0.584	0.698	0.280

The results show that the web-based results are sensibly worse than those obtained on the static collection. This is due primarily to two reasons: in the first place, because topics were tailored on the GeoCLEF collection. Therefore, some topics refer explicitly to events that are particularly relevant in the collection and are easier to retrieve. For instance, query `GC-005` "Japanese Rice Imports" targets documents regarding the opening of the Japanese rice market for the first time to other countries; "Japan" and "Rice" in the document collection appear together only in such documents, therefore it is easier to retrieve the relevant documents when searching the GeoCLEF collection.

The second factor affecting the results for the Web-based system is the ambiguity of toponyms, which does not allow to correctly estimate the probabilities for places. For instance, in the results obtained for topic `GC-008` ("Milk Consumption in Europe"), the MI obtained for "Turkey" was abnormally high with respect to the expected value for this country. The reason is that in most documents, the name "turkey" was referring to the animal and not to the country. This kind of ambiguity represents one of the most important issue at the time of estimating the probability of occurrence of places. Ambiguity (or, better, the polysemy of toponyms) grows together with the size and the scope of the collection being searched. The GeoCLEF collection was also semantically tagged using WordNet and Geonames IDs to identify the places referenced by toponyms, while Web content is rarely tagged using precise IDs, therefore increasing the chance of error in the estimation of probabilities for places which share the same name.

There are three kind of toponym ambiguity that can be recognised (after the two main types identified by [11]:

- Geo / Non-Geo ambiguity: in this case, a toponym is ambiguous with respect to another class of name (such as "Turkey" which may be the animal or the country);
- Geo / Geo ambiguity of different class: for instance, "Puebla" the city or the state;
- Same class Geo / Geo ambiguity.

The solution in all cases would be to use an ontology to precisely identify places in documents; the only difference is the amount of information that the ontology should include. For the first type of ambiguity, the only information needed is whether the name represents a place or not. In the second case, we would also need to know the class of the place. Finally, in the Geo / Geo ambiguity, we may differentiates places using their coordinates or by knowing the including entity, or both. The Geonames ontology contains all these information and represents the best option at the time of geographically tag place names.

4 Conclusions

The results obtained with Geooreka! over a static, semantically-labelled (at least from a geographical viewpoint) collection compared to the results obtained in

the Web showed that the imprecise identification of places is a problem for search engines destined to users who are interested in searching for geographically constrained information. The use of precise semantically tagging schemes for toponyms, such as Geonames RDF, would allow these search engines to produce more reliable results. Spreading the use of geographical tagging for the Semantic Web would also allow users to mine information using geographical constraints in a more effective way. In this sense, we would like to encourage the use of Geonamen in order to produce accurate geographically tagged Web content.

References

1. Sanderson, M., Kohler, J.: Analyzing geographic queries. In: Proceedings of Workshop on Geographic Information Retrieval (GIR04). (2004)
2. Gan, Q., Attenberg, J., Markowetz, A., Suel, T.: Analysis of geographic queries in a search engine log. In: LOCWEB '08: Proceedings of the first international workshop on Location and the web, New York, NY, USA, ACM (2008) 49–56
3. Andogah, G.: Geographically Constrained Information Retrieval. PhD thesis, University of Groningen (2010)
4. Boll, S., Jones, C., Kansa, E., Kishor, P., Naaman, M., Purves, R., Scharl, A., Wilde, E.: Location and the web (locweb 2008). In: Proceeding of the 17th international conference on World Wide Web. WWW '08, New York, NY, USA, ACM (2008) 1261–1262
5. Buscaldi, D., Rosso, P.: Geooreka: Enhancing Web Searches with Geographical Information. In: Proc. Italian Symposium on Advanced Database Systems SEBD-2009, Camogli, Italy (2009) 205–212
6. Buscaldi, D., Rosso, P., Sanchis, E.: Using the WordNet Ontology in the GeoCLEF Geographical Information Retrieval Task. In Peters, C., Gey, F.C., Gonzalo, J., Mller, H., Jones, G.J., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D., eds.: Accessing Multilingual Information Repositories. Volume 4022 of Lecture Notes in Computer Science. Springer, Berlin (2006) 939–946
7. Buscaldi, D., Rosso, P.: On the relative importance of toponyms in geoclef. In: Advances in Multilingual and Multimodal Information Retrieval, 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers, Springer (2007) 815–822
8. Giuliano, C.: jWeb1T: a library for searching the Web 1T 5-gram corpus. (2007) Software available at http://tcc.itc.it/research/textec/tools-resources/jweb1t.html.
9. Kullback, S., Leibler, R.A.: On Information and Sufficiency. Annals of Mathematical Statistics **22**(1) (1951) pp. 79–86
10. Buscaldi, D., Rosso, P.: Using GeoWordNet for Geographical Information Retrieval. In: Evaluating Systems for Multilingual and Multimodal Information Access, 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers. (2009) 863–866
11. Amitay, E., Harel, N., Sivan, R., Soffer, A.: Web-a-where: Geotagging web content. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK (2004) 273–280

SV: a visualization mechanism for ontologies of records based on SVG graphics

Ma. Auxilio Medina, Miriam Cruz, Rebeca Rodríguez, Argelia B. Urbina

Universidad Politécnica de Puebla
Tercer Carril del Ejido Serrano S/N
Juan C. Bonilla, Puebla, México
{mmedina, mcruz, rrodriguez, aurbina} @uppuebla.edu.mx,
WWW home page: http://informatica.uppuebla.edu.mx/
~mmedina, ~rrodriguez, ~aurbina

Abstract. This paper describes SV, a visualization mechanism used to explore digital collections represented as hierarchical structures called ontologies of records. These ontologies are XML files constructed using OAI-PMH records and a clustering algorithm. SV is composed by a web interface and SVG graphics. Through the interface, users can recognize the organization of the collection and access to metadata of documents.

1 Introduction

Digital libraries gather valuable information. Organizations such as the Open Archives Initiative (OAI[1]) have proposed different alternatives to share data. The Protocol for Metadata Harvesting (OAI-PMH protocol), for example, supports interoperability between federated digital libraries. Documents are described in metadata records. Dublin Core Metadata (DC[2]) is the default metadata format for this protocol.

The services and the collections of digital libraries are enriched in the Semantic Web. The use of XML, Resource Description Framework (RDF), OWL, conceptual maps and other metadata technologies are addressed to improve search tasks [1]. Semantic Digital Libraries (SDLs) refer to systems build upon digital libraries and social networking technologies (Web 2.0) [2]. Freely distributed software exists to construct SDLs such as Greenstone[3] or Jerome DL[4]. In this type of software, ontologies play a key role, they refer to explicit specifications of shared conceptualizations [3]. Ontologies enables the representation of knowledge that software and human agents can understand and use.

This paper proposes the use of ontologies called "ontologies of records" that are represented as XML documents as the basis of a visualization mechanism

[1] http://www.openarchives.org/
[2] http://dublincore.org
[3] http://www.greenstone.org/
[4] http://www.jeromedl.org/

called *semantic view (SV)*. The name also refers to the first two letters of "Support Vector Graphics". SV offers an interactive view to allow users to explore the content of a federated collection.

The paper is organized as follows. Section 2 describes the features of an ontology of records. Section 3 includes related work. Section 4 and 5 explains the design and implementation of SV, respectively. Experimental results are described in Section 6. Finally, Section 7 includes conclusions and suggests future directions of our work.

2 What is an ontology of records

An *ontology of records* is a hierarchical structure of clusters of OAI-PMH records that provides an unambiguous interpretation of its elements. Its construction is based on the *Frequent Itemset Hierarchical Clustering* algorithm [8]. This structure organizes a collection of documents, this has concept-term relationships useful for keyword based searches. An ontology of records is stored as a well formed XML file that is validated against an XML Schema. An ontology of records has the following features[9]:

1. Documents are clustered by similarity
2. Clusters in the k-level have labels of k-terms
3. All the records of a cluster share the terms of its label

3 Related work

This section describes some systems that have been used to visualize collections of documents. Proat et al. [4] use 3D trees to visualize documents organized according to the Library of Congress Classification (LCC). Documents are clustered in seven subsets. The interface has controls to rotate or zoom the nodes of trees. The leaf nodes contain metadata of documents.

Geroimenko et al. [5] have proposed the Generalizad Document Object Model tree Interface (G-DOM-Tree interface) to visualize metadata from XML DOM (Document Object Model) documents. The model displays a hierarchy of labels, this is very similar to the visualization that browsers offer of XML Schema. The interface is implemented as a Java applet or a Flash film.

Fluit et al. [6] describe *Spectacle*, this mechanism uses lightweight ontologies to represent classes of similar objects and their relationships. The navigation can be done by using hypertext or "cluster maps". A cluster map visualizes the objects and their classes.

At last, Sánchez et al. [7] use a star field grid to visualize documents from several collections. Documents are stored as OAI-PMH records. The axis of the grid represent attributes of the collections that can be chosen by users. Small polygons are associated with the type of document and different colors are used to distinguish the collections.

4 Desing of SV

The design of SV is addressed to reach the following objectives:

- Construct a visualization mechanism with semantic features that allow users to explore a collection of documents
- Represent the organization of a collection of documents
- Retrieve the metadata and the content of a determined document

In order to reach these objectives, we have used the levels of knowledge proposed by [2] in the design of SV. We want to uses CORTUPP as a test bed, this is a collection represented as an ontology of records[5].

1. **Level 1: Organization of the metadata**. Metadata is organized in the ontology of records. Content information is stored in `dc:title`, `dc:subject` and `dc:description` elements.
2. **Level 2: Organization of the information in the documents**. Technical reports have a common structure formed by six mandatory chapters: 1)research propose, 2)state of the art, 3)research design, 4)implementation, 5)results and 6)conclusions. This structure is defined in a Latex template. The BibTex file format is used to manage the bibliography. A technical report is described as a `@techreport` entry.
3. **Level 3: Organization of the information in databases**. The technical reports are stored as PDF files in a database that also includes data and counts of users. Documents are accessible through a web interface.
4. **Level 4: Organization of the topics treated in the documents**. The `dc:subject` element stores the topic of a document. Keywords of this element belong to the labels of the clusters in the ontology of records.
5. **Level 5: Organization of the concepts, terms and relations**. This level is also represented in the ontology of records.

5 Implementation of SV

SV is formed by a web interface and SVG graphics[6]. SVG is a format developed and maintained by the W3C SVG Working Group. This is an XML application used to describe animated or static two dimensional vectorial graphics. The main feature of these graphics is scalability.

SV uses Xerces, this is a Java parser used to extract data from an ontology of records. The classes of SV are built using Java language. In the interface, each document, that is, an OAI-PMH record, is represented with a yellow star in a blue gradient background. The background is divided in five parts that correspond to the first levels of the ontology. These levels are divided by lines that form angles of 90 degrees. The distribution of the lines try to reflect an estimation of the amount of documents that can be found in each level. The documents closer to

[5] CORTUPP is available at http://server3.uppuebla.edu.mx/cortupp/
[6] http://www.w3.org/svg/

the upper left corner belong to the first level of the ontology, these documents share one term. The second level shows the documents that share two terms, and then on. The stars have different size according to their level, they are bigger at the first level and smaller at the last one.

The interface of SV is a SVG graphic of 502 per 502 pixels. XML Parser is the Java application used to construct the XML document that contains the interface. XLink is used to create hyperlinks between documents and their metadata. Given a click on a star, users can allow the metadata on the right panel. Figure 1 shows the SV interface where only six documents at the second and third level were included, however SV is designed to support until 500 documents. The colors can be modified without requiring compilation because they are stored in a text file. The mechanism is accessible at http://informatica.uppuebla.edu.mx/ visualizacionPI/index.html.

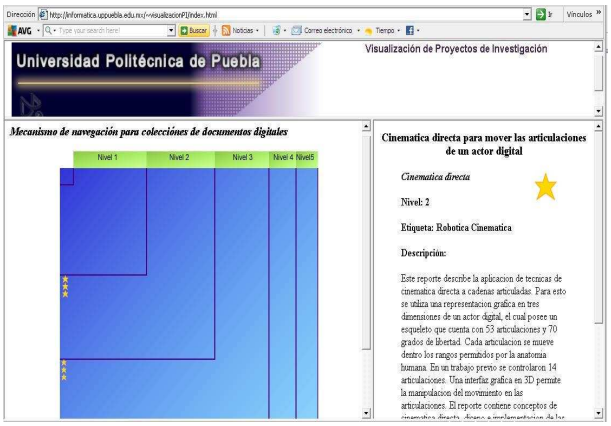

Fig. 1. Using VS to visualize CORTUPP

6 Experimental results

Different configuration of ontologies of records were constructed in order to check SV, that is, unity tests and integration tests were performed successfully. After the installation of the SVG Plugin Version 1.7, the visualization of SV was successful using Internet Explorer 8, Google Chrome 7.0.517.41 and Opera 10.6, however, there were some inconveniences using Firefox 1.5, Firefox 3.6 and Firefox Beta due to these versions do not support the animation features of SVG graphics.

7 Conclusions

We have described SV, a visualization mechanism of federated collections based on ontologies. SV has semantic features represented in the interface such as the location of documents in the ontology and the similarity between documents. Additional semantic information is stored in the metadata attached to each document and in the ontology of records. Through SV interface, users can access to metadata or download a document.

CORTUPP was used as a test bed for SV, however, any collection of OAI-PMH records represented as an ontology of records can be visualized. Although the size of an ontology of records can impact the visualization of SV, its design is flexible enough to support distinct collections. As future work, we plan to expand SV to show the clusters and their labels. Then, we would like to incorporate tagging and recommendation mechanisms.

References

1. Geroimenko V., C.C.: Visualizing the Semantic Web. XML-based Internet and Information Visualization. Segunda edición edn. Volume 1. Springer, Wokingham, England (2003) Libro. los primeros cuatro datos son obligatorios.
2. Kruk, S.R., McDaniel, B.: Semantic Digital Libraries. Springer-Verlag, Berlin, Heidelberg (2009)
3. Gruber, T.: A translation approach to portable ontology specification. Knowledge Adquisition **5**(2) (1993) 199–220
4. C., P.: Sistema uva: interfaces para visualización de grandes colecciones digitales. Tesis de maestría, Universidad de las Américas Puebla, Santa Catarina Mártir S/N, San Andrés Cholula, Puebla, México. (2002) Tesis de maestría. Los primeros cuatro campos son obligatorios.
5. Geroimenko V., G.L.: Interactive interfaces for mapping e-commerce ontologies in Visualizing the Semantic Web. XML-based Internet and Information Visualization. Segunda edición edn. Volume 1. Springer, Wokingham, England (2003) Libro. los primeros cuatro datos son obligatorios.
6. Fluit C., Sabou N., H.v.F.: Ontology-based information visualization. In: Visualizing the semantic web: XML based internet and information visualization. Volume 1. Segunda edición edn. Springer, Wokingham, England (2002)
7. Sánchez J. A., Quintana M. G., R.A.: Star-fish: Starfields+fisheye visualization and its application to federated digital libraries. Proceedings of the 3rd Latin American Conference on Human-Computer Interaction (CLIHC 2007, Nov.) (2007)
8. Fung, B., Wang, K., Ester, M.: Hierarchical document clustering using frequent itemsets. In: Proceedings of the Third SIAM International Conference on Data Mining, (SDM'03, San Francisco, California, May),, San Francisco, CA, USA, SIAM (2003) 59–70
9. Medina, M.A., Sánchez, J.A.: Ontoair: A method to construct lightweight ontologies from document collections. Mexican International Conference on Computer Science **0** (2008) 115–125

Modeling of CSCW system with Ontologies

Mario Anzures-García[1, 2], Luz A. Sánchez-Gálvez[1,2], Miguel J. Hornos[2], Patricia Paderewski-Rodríguez[2], and Antonio Cid[1]
[1] Facultad de Ciencias de la Computación, Benemérita Universidad Autónoma de Puebla, 14 sur y avenida San Claudio. Ciudad Universitaria, San Manuel, 72570 Puebla, Mexico
{anzures, luzsg}@correo.ugr.es
[2] Departamento de Lenguajes y Sistemas Informáticos, E.T.S.I. Informática y de Telecomunicación, Universidad de Granada, C/ Periodista Saucedo Aranda, s/n, 18071 Granada, Spain.
{mhornos, patricia}@ugr.es

Abstract. In recent years, there has been a growing interest in the development and use of domain ontologies, strongly motivated by the Semantic Web initiative. However, the application of ontologies in the CSCW domain has been scarce. Therefore in this paper, it presents a novel architectural model to CSCW systems described by means of an ontology. This ontology defines the fundamental organization of a CSCW system, represented in its concepts, relations, axioms and instances.

Keywords: Ontology, Groupware Application, SOA, Architectural Model, Services.

1 Introduction

In the last two decades, the enormous growth of Internet and the web have given rise to an intercreativity cyberspace, in which groups of people can communicate, collaborate and coordinate to carry out common tasks. Therefore, a great number of groupware applications has been developed using different approaches, including object-oriented, component-oriented, and agent-oriented ones. However, the development of this kind of applications is very complex, because different elements and aspects must be taken into account. Hence, these applications must be simultaneously supported by models, methodologies, architectures and platforms to be developed in keeping with current needs. In the groupware domain, one of the models most used is the Unified Modelling Language (UML) [1], although this has not any element to represent constrains, which are very important in applications so complex as the groupware ones.

There has recently been an increase in the use of ontologies in any domain to model applications. An ontology is presented as an organization resource and knowledge representation through an abstract model. This representation model provides a common vocabulary of a domain and defines the meaning of the terms and the relations amongst them. In the domain of groupware applications, the ontology

provides a well-defined common and shared vocabulary, which supplies a set of concepts, relations and axioms to describe this domain in a formal way.

In this paper, two ontologies for the groupware domain are proposed. The first ontology determines who authorize the registration of users, how interaction is carried out among them, and how the turns for users participation are defined, among other aspects. Moreover, it allows supporting modifications in runtime, such as changing the user role, the rights/obligations of a role, the current policy, etc. The second ontology establishes the necessary SOA-based services to develop groupware applications in accordance with the existing papers in the literature about the development of this type of applications. In addition, these services are clustered in modules and layers with respect to the concern that they represent.

This paper is organized as follows. Section 2 gives an brief introduction to the ontologies. Section 3 describes the ontology-based modeling of the group organizational structure. Section 4 presents an ontological model, which allows us to specify an architectural model for the development of groupware applications. Finally, Section 5 outlines some conclusions and future work.

2 Introduction to the Ontologies

There are several definitions of ontology, which have different connotations depending on the specific domain. In this paper, we will refer to Gruber's well-know definition [2], where an ontology is an explicit specification of a conceptualization. For Gruber, a conceptualization is an abstract and simplified view of the world that we wish to represent for some purpose, by the objects, concepts, and other entities that are presumed to exist in some area of interest, and the relationships that hold them. Furthermore, an explicit specification means that concepts and relations need to be couched by means of explicit names and definitions.

Jasper and Ushold [3] identify four main categories of ontology applications: 1) neutral authoring, 2) ontology-based specification, 3) common access to information, and 4) ontology-based search. In the work presented here, the main idea is to use ontologies to specify the modeling of both the group organizational structure and the architectural model in the groupware domain, since an ontology is a high level formal specification of a certain knowledge domain, which provides a simplified and well defined view of such domain.

Ontology is specified using the following components:

- *Classes*: There is a set of classes, which represent concepts that belong to the ontology. Each class may contain individuals (or instances), other classes or a combination of both, with their corresponding attributes.
- *Relations*: These define interactions between two or several classes (object properties) or between a concept and a data type (data type properties).
- *Axioms*: These are used to impose constraints on the values of classes or instances. Axioms represent expressions (logical statement) in the ontology and are always true inside the ontology.
- *Instances*: These represent the objects, elements or individuals of an ontology.

These four components will be described for the two ontologies proposed in this paper.

In addition, ontologies require of a logical and formal language to be expressed. In Artificial Intelligence, different languages have been developed, like the First-Order Logic-based (which provide powerful primitive for modeling), the Frames-based (with more expressive power but less inference capacity), and the Description Logics-based (which are more robust in the reasoning power) ones.

OWL (Web Ontology Language) [4] is a language based on Description Logics for defining and instantiating Web ontologies based on XML (eXtensible Markup Language) [5] and RDF (Resource Description Framework) [6]. OWL can be used to explicitly represent the meaning of terms in vocabularies and the relationships among those terms. This language makes possible to infer new knowledge from a conceptualization, by using a specific software called *reasoner*. It has used the tool Protégé [7], which is based on OWL, to define the ontology for group organizational structure.

In the groupware domain, ontologies have mainly been used to model task analysis or sessions. Different concepts and terms, such as group, role, actor, task, etc. have been used for the design of task analysis and sessions. Many of these terms are considered in our conceptual model. Moreover, semiformal methods (e.g. UML class diagrams, use cases, activity graphs, transition graphs, etc.) and formal ones (such as algebraic expressions) have also been applied to model the sessions. There is also a work [8] for modeling cross-enterprise business processes from the perspective of cooperative system, which is a multi-level design scheme for the construction of cooperative system ontologies. This last work is focused on business processes, and it describes a general scheme for the construction of ontologies. However, in this paper, we propose to model two specific aspects: the group organizational structure and the architecture of a groupware application. Consequently, the application domain of both ontologies is groupware, not business processes.

3 Ontology for specifying an architectural model

In order to specify architectural model five concerns are identified: Data, Group, Cooperation, Application, and Adaptation. Consequently, five layers are considered. Four layers are composed by modules and services, while the fifth one, the *Data Layer*, contains repositories with the necessary information to carry out the group work. The services of the architectural model are defined by the concepts' ontology.

3.1. Ontology Concepts

The architecture components are characterized through the concepts' ontology (shown in Figure 1), which will be briefly described below:

- *Registration* is the first action that a user must carry out to can participate in the group work using the collaborative application.
- *Authentication* validates the access to the group and depends on the organizational style defined in the same.

- *Group* is who works in the session to perform work group.
- *Organizational_Style* defines the organizational style that a group will use to carry out the group work.
- *Stage* restricts user's access to the application in accordance with the organizational style defined in it.
- *Session* defines a shared workspace where a group carries out common tasks.
- *Session_Management* manages and controls one or more sessions.
- *Concurrency* manages shared resources to avoid inconsistencies by using them.
- *Shared_Resource* is used by users to carry out basic activities.
- *Basic_Activity* is an action that a user must perform to carry out a task (which can be made up by one or more basic activities).
- *Task* is carried out by the group to achieve a common goal.
- *Notification* notifies one or more users of all events that happen in a session.
- *Group_Awareness* gets the necessary information to supply group awareness to users that take part in a group.
- *Group_Memory* is supplied by the application to facilitate a common context.
- *Application* is used by the users to carry out group work in established session.
- *Configuration* configures the application the first time that it is used and when it is necessary.
- *User_Interface* shows users all the information about the application execution.
- *Environment* modifies the user interface to present the information in accordance with the device used by each user.
- *Adaptation* is a process that allows adapting the collaborative application to the new needs of the group.
- *Detection* monitors the execution environment to detect the events that determine the adaptation process.
- *Agreement* decides whether an adaptation process must be carried out or not.
- *Vote_Tool* is used by users to perform the agreement.
- *Adaptation_Flow* is a set of steps carried out to adapt the collaborative application in accordance with the selected event.
- *Repair* is required when the adaptation process can not be performed.

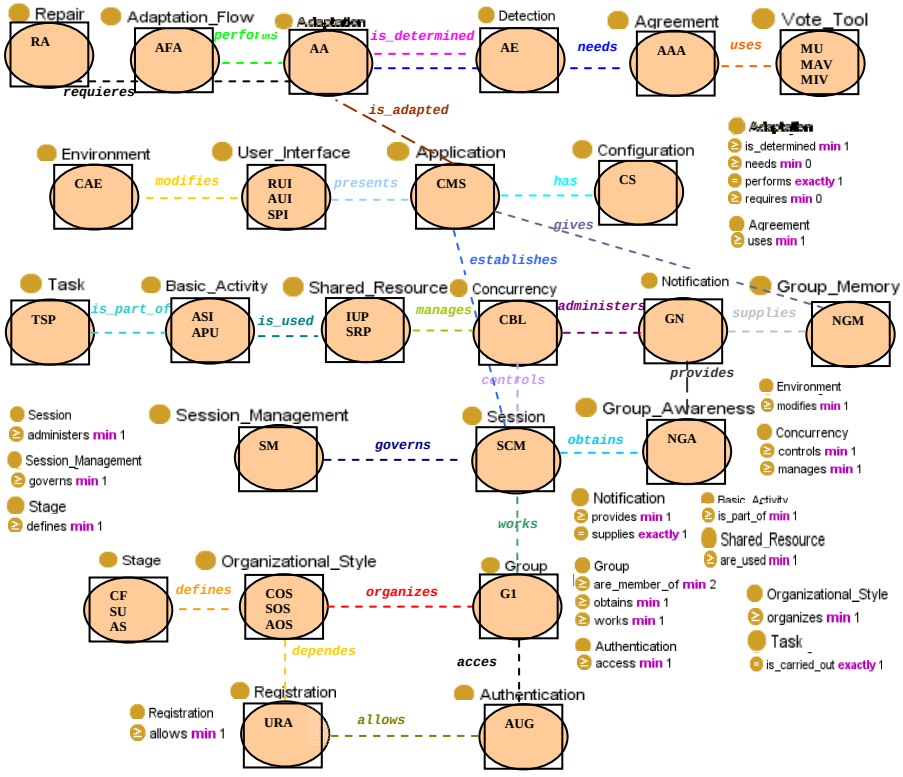

Figure 1. Ontology for specifying an architecture to model collaborative applications.

3.2. Ontology Relations

The architecture relationship to each component and its environment are symbolized with the ontology relations (see Figure 1) listed below:

- *allows (Registration, Authentication):* Only registered users are allowed to authenticate to access to the collaborative application.
- *access* (*Authentication, Group):* Authentication allows users to access to group.
- *depends* (*Registration, Organizational_Style):* Users registration depends on the organizational style defined at a given stage.
- *organizes (Organizational_Style, Group):* An organizational style specifies the way in which the group is organized.
- *defines (Stage, Organizational_Style):* A stage defines an organizational style.
- *works (Group, Session):* A group needs to be connected to a session to work.
- *governs (Session_Management, Session):* The session management governs a session.
- *controls (Concurrency, Session):* The concurrency service controls the existing interaction in a session.
- *manages (Concurrency, Shared_Resource):* The concurrency service manages the shared resources to guarantee mutually exclusive usage of these.
- *is_used (Shared_Resource, Basic_Activity):* The shared resources are used by basic activities.
- *is_part_of (Basic_Activity, Task):* A basic activity is part of a task.
- *administers (Session, Notification):* The session administers the notification.
- *provides (Notification, Group_Awareness):* The notification process provides group awareness.
- *obtains (Group, Group_Awareness):* A group obtains group awareness to avoid inconsistencies in the collaborative application.
- *supplies (Notification, Group_Memory):* The notification process supplies group memory.
- *gives (Application, Group_Memory):* The application gives group memory.
- *establishes (Application, Session):* An application establishes a session.
- *presents (Application, User_Interface):* An application presents an user interface so that users can use the collaborative application.

- *modifies (Environment, User_Interface):* The environment modifies the user interface according to the device used by each user.
- *has (Application, Configuration):* Each application has a configuration process, which is carried out by users.
- *is_adapted (Application, Adaptation):* An application is adapted by the adaptation process.
- *is_determined (Adaptation, Detection):* The adaptation process is determined by the detection process.
- *needs (Adaptation, Agreement):* The adaptation process needs an agreement process to decide whether the adaptation is carried out or not.
- *uses (Agreement, Vote_Tool):* The agreement process uses a vote tool to carry out the agreement.
- *performs (Adaptation, Adaptation_Flow):* The adaptation process performs an adaptation flow to appropriately adjust the application.
- *requires (Adaptation, Repair):* When the adaptation process can not be performed, it is required to repair the application to avoid inconsistencies in it.

3.3. Ontology Axioms

Finally, the principles governing design and evolution of the architectural model are represented by ontology axioms (see Figure 1):

- An authentication must have only one registration, i.e. an user is authenticated only if she/he is registered.
- A registration depends on an organizational style, i.e. an user is registered with accordance to organizational style established in the group work.
- An organizational style organizes at least one group.
- A group works at least in one session.
- An application establishes at least one session.
- A session administers at least one notification process.
- A group obtains group awareness.
- An application gives group memory.
- The concurrency service controls at least one session.
- The concurrency service manages at least one shared resource.
- A shared resource is used by at least one basic activity.
- A basic activity is part of at least one task.
- An application has at least one possible configuration.
- An application presents at least one user interface.
- An environment modifies at least one user interface.
- An application can be adapted by an adaptation process.
- An adaptation process is determined by at least one detection process.
- An agreement process is carried out only if there is an adaptation process in a non-hierarchical organizational style.
- An agreement process uses at least one vote tool.
- An adaptation process performs only one adaptation flow.
- An adaptation flow must verify at least one pre-condition and post-condition to carry out the adaptation.

- An adaptation process can require a repair process, if this has not finished.

3.4. Ontology Instances

In order to show the architectural functionality, this section presents a set of instances (see Figure 1), derived from the definition of the application instance, which is a Conference Management System (CMS). A CMS is a web-based application that supports the organization of scientific conferences. It can be regarded as a domain-specific content management system. Nowadays, similar systems are used by editors of scientific journals.

This type of systems generally has four stages: submission, assignment, review, and acceptance/rejection of papers, and this paper adds the stage of application configuration. CMS supports three user groups: Authors (A), Program Committee Members (PCM) and Program Committee Chairs (PCC). The first user group (A) corresponds to people who can submit papers (at the submission stage) through the Internet, and who receive the review results and the final decision via an email (at the acceptance/rejection stage). It is the largest user group (its average number is normally between 100 and 400 people for most conferences). The second user group (PCM) is made up of people who must evaluate some of the submitted papers and send the result to the PC Chairs via the Internet (at the review stage). Its number is about 20–50 persons in average. People in the last user group (PCC) are in charge of allocating papers to reviewers (at the assignment stage) and making the final decision on papers, as well as a number of other operations. This is the least numerous group, being usual 1–3 PCC per conference. Therefore:

- *Session* instance is Session of the Conference Management (SCM).
- *Session_Management* instance is Session Management (SM).
- *Stage* instances are configuration (CF), and submission (SU), assignment (AS), review (RE) and acceptance/rejection (AC) of papers.
- *Authentication* instance is user authentication in the group (UAG).
- *Registration* instance is user registration in the application (URA).
- *Users* instances are U1, U3 and U4 as A, U3 as PCM, and U2 as PCC.
- *User_Interface* (UI) instances are Registration UI (RUI), Authentication UI (AUI), Submitting Paper UI (SPI), Configuration UI (CUI), etc.
- *Environment* instance is collaborative application environment (AE).
- *Organizational_Style* (OS) instances are Configuration OS (COS), Submission OS (SOS), Assignment OS (AOS), Review OS (ROS), and Acceptance/rejection OS (POS). In the ontology shown in Figure 3, SOS is the unique OS considered by simplicity reasons.
- *Group* instance is G1, which is made up of three users, U1, U3 and U4, because U2 does not participate at SOS.
- *Concurrency* instance is locks mechanism (LM).
- *Shared_Resource* (SR) instances are paper (SRP), and uploading paper (IUP).
- *Basic Activity* (BA) instances are submitting information (ASI), and uploading paper (AUP).
- *Task* instance is submitting paper (TSP).
- *Notification* instance is group notification (GN).

- *Configuration* instance is configuration of the system (CS).
- *Adaptation* instance is adaptation application (AA).
- *Detection* instance is adaptation event (AE).
- *Agreement* instance is adaptation agreement of the application (AAA).
- *Vote Tool* instances are majority vote (MV), maximum value (MAV) and minimum value (MIV).
- *Adaptation_Flow* instance is adaptation flow of the application (AFA).
- *Repair* instance is reparation of the adaptation (RA).

4 BPM to Manage the Ontoloy-based Architectural Model

BPM [16] is a set of methods, tools and technologies used to design, perform, analyze and manage operational business processes, by means of different phases. It also facilitates service composition. In this paper, BPM is based on ontological approach [17] and is composes for three phases, which are: *Process Modeling*, *Process Implementation*, and *Process Execution*. The ontology is used in order to simplify the task of governing the behaviour of BPM; it enables to BPM to uses concepts to describe the models and the entities being controlled, thus simplifying their description and facilitating the analysis and the careful reasoning over them; and it allows dynamically calculating relations between business processes and environment, supporting modifications in runtime. BPM and SOA make the integration faster and easier than never, it is not necessary to discard of the investments already made, everything can be reused.

4.1. Process Modeling

This phase identifies the participating elements in a business process. Unlike other existent models, we use an ontology (shown in the Figure 1), to clearly specify the semantics of the tasks and the decisions in the process flow. Therefore, four layers of the architectural model, (see Figure 2), are composed of the services, which were defined as concepts, in the ontology. The *Group Layer* includes three modules, which are *Access, Group* and *Session*. The *Access Module* has two services: *Registration* and *Authentication*. The *Group Module* presents three services: *Group, Organizational Style,* and *Stage*. The *Session Module* contains two services: S*ession Management,* and *Session.* The *Cooperation Layer* has the *Context Module* and *Interaction Module*. The former includes four services: *Concurrency, Shared Resource, Activity Basic,* and *Task.* The later encompasses the *Group Awareness Service,* the *Group Memory Service,* and the *Notification Service.* The *Application Layer* comprises only the *Application Service,* the *Configuration Service,* the *User Interface Service,* and the *Environment Service.* Finally, the *Adaptation Layer* involves the *Pre-adaptation Module* and the *Adaptation Module.* The former encompass the *Detection Service,* the *Agreement Service,* and the *Vote Tool Service.* The latter comprises the *Adaptation Flow Service,* and the *Repair Service.*

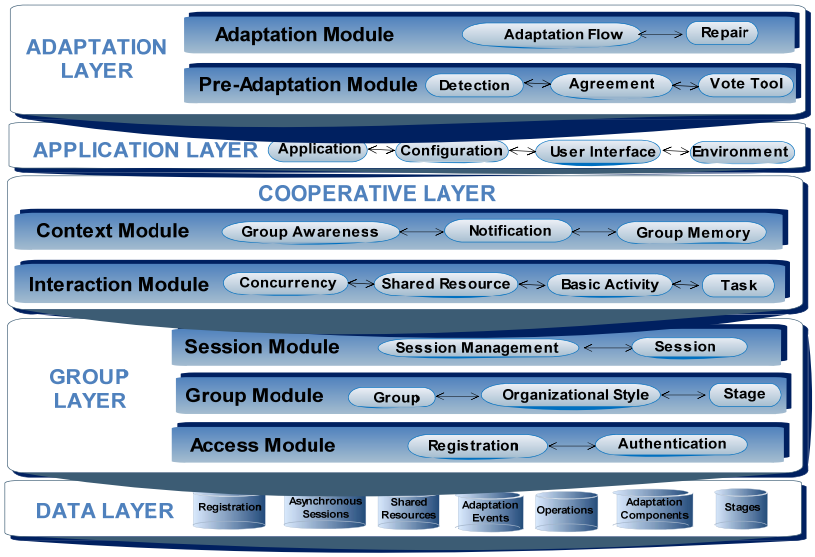

Figure 2. Architectural model for developing collaborative applications.

4.2. Process Execution

In this phase, the business process model is transformed into an executable process model, which can be deployed to a process engine for its execution. Figure 3 shows a sequence diagrams (that represents an executable process model), when the author submits papers to the CMS. In this figure, the CMS users are consumer services, invoking different services and only are considering some services of the architectural model by simplicity.

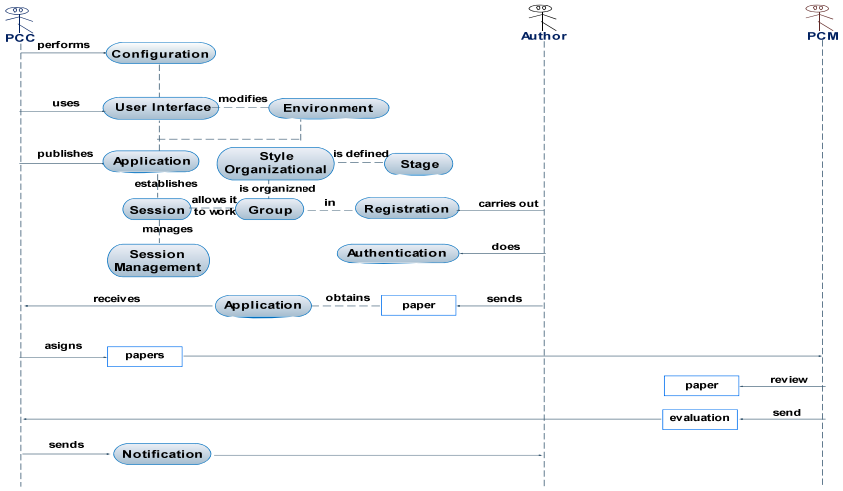

Figure 3. Sequence diagram of papers sent.

4.3. Process Implementation

In the process execution phase, the process engine executes a process model by firstly creating a process instance and then navigating through the control flow of the process model. In order to ensure seamless interaction when navigating through the control flow of the process model, this phase provides mechanisms for the discovery, selection and invocation of services. The module dynamically discovers and selects the appropriate service of the architectural model basing on the task description, and invokes it on behalf of the process engine, which plays the role of a *requester service* when it invokes the service to perform a task. Moreover, this module carries out a process monitoring that provides relevant information about running process instances. If during the process execution a failure arises, such as network faults, server crashes, or application-related errors (e.g. unavailability of a requested service, errors in the composition or missing data, etc.), reconfiguration actions are carried out, such as duplication (or replication) or substitution of a faulty service. The first case involves addition of services representing similar functionalities; this aims at improving load balancing between services in order to achieve a better adaptation. The second case encompasses redirection between two services; applying this action means the first one is deactivated and replaced by the second one.

5 Conclusions and Future Work

The current work has presented an ontology-based architectural model, which facilitates the development of collaborative applications. The ontology describes the components, their relationships to each other and the environment, and the principles governing architectural design and evolution. For that reason, we think that the ontology is a proper model to describe architectures. BPM is used to manage and control the interaction between the services that make up the architectural model. In addition, BPM also is based on the ontology proposed. These services are designed using SOA that together with BPM, facilitates the application's integration. The future work will consist on extending the existent reconfiguration actions of the service-based collaborative applications.

References

1. Garlan D., Shaw, M.: An introduction to software architecture. Advances in Software Engineering and Knowledge Engineering, 1--39, (1993)
2. Perry, D.E., Wolf, A.L.: Foundations for the study of software architecture. ACM SIGSOFT Software Engineering Notes 17(4), 40--52, (1992)
3. Architecture working group: Recommended practice for architectural description of software-intensive systems. IEEE Std 1471 (2000)
4. UML 2.0 Superstructure Specification (OMG). Ptc/03-08-02, 455—510, (2003)
5. Spivey, J.M.: The Z Notation: A Reference Manua., Prentice Hall, (1989)
6. Abrial, J.R.: The B-book: Assigning Programs to Meanings. Cambridge University Press, (1996)

7. Gruber, T.R.: Toward Principles for the Design of Ontologies Used for Knowledge Sharing. I. J. Human Computer Studies 43-(5/6), 907--928 (1995)
8. Gómez-Pérez, A., Fernández-López, M, Corcho, O.: Ontological Engineering with Examples from the Areas of Knowledge Management, e-Commerce and the Semantic Web, Springer, (2004)
9. Uschold, M., Grüninger, M.: Ontologies: Principles, Methods and Applications. Knowledge Engineering Review 11(2), 93–155 (1996)
10. Farquhar, A., Fikes, R, Rice, J.: The Ontolingua Server: A Tool for Collaborative Ontology Construction. I. J. Human Computer Studies 46(6), 707–727, (1997)
11. Dean, M., Schreiber, G.: OWL Web Ontology Language Reference. W3C Working Draft. http://www.w3.org/TR/owl-ref/ (2003)
12. Protegé: http://protege.stanford.edu/
13. Noguera, M., Hurtado, V, Garrido, J.L.: An Ontology-Based Scheme Enabling the Modeling of Cooperation in Business Processes. In; Meersman, R., Tari, Z., Herrero, P. (eds.) OTM Workshops 2006. LNCS vol. 4277, pp. 863—872. Springer, Heidelberg (2006)
14. Erl, T.: SOA: Concepts, Technology and Design. Prentice-Hall, (2005)
15. Howard, S., Fingar, P.: Business Process Management: The Third Wave, Meghan-Kiffer, (2003)
16. May, M.: Business Process Management: Integration in a Web-Enabled Environment, Prentice Hall, (2003)
17. Hepp, M., Roman, D.: An Ontology Framework for Semantic Business Process Management, In: 8th International Conference on Wirtschafts Informatik, Vol. 1 pp. 42--440, (2007)

The use of WAP Technology in Question Answering

Fernando Zacarías F.[1], Alberto Tellez V.[2], Marco Antonio Balderas[3], Guillermo De Ita L., and Barbara Sánchez R.[4]

Benemérita Universidad Autónoma de Puebla,
[1,3,4,5]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]brinza@hotmail.com

Abstract. The experience of Puebla Autonomous University on using WAP technology in the development of novel applications is deployed. The goal is to enhance question answering through innovative mobile applications providing new services and more efficiently. The architecture proposed based on WAP protocol, 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. This characteristic gives to Question Answering an added value. Furthermore, the method for answering definition questions is very precise. It could answer almost 95% 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 a new application based on Wikipedia that makes Question Answering 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 service in anytime, anywhere and any device (PDA's, cell-phone, NDS, etc.). Furthermore, our application can be accessed via Web through iPhone and any device with internet access.

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, we present the method for question answering for definitions questions in section 3. After, in section 4 we present the WAP technology as support for our mobile application. Section 5 shows our application on the two variants, WiFi and WAP protocol. Section 6 describe our perspectives about our future work. Finally, the conclusions are drawn in section 7.

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 an-

swering 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:

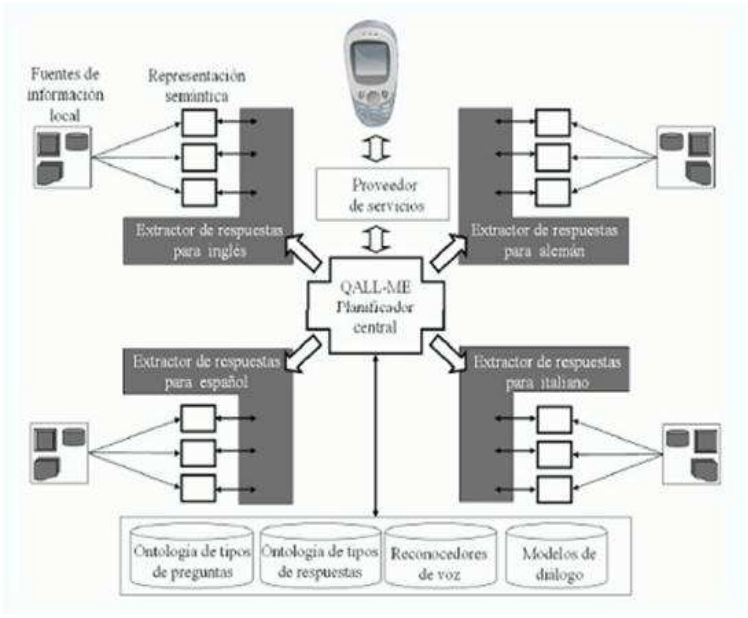

Fig. 1. Main QALL-ME Architecture [8]

"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 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.

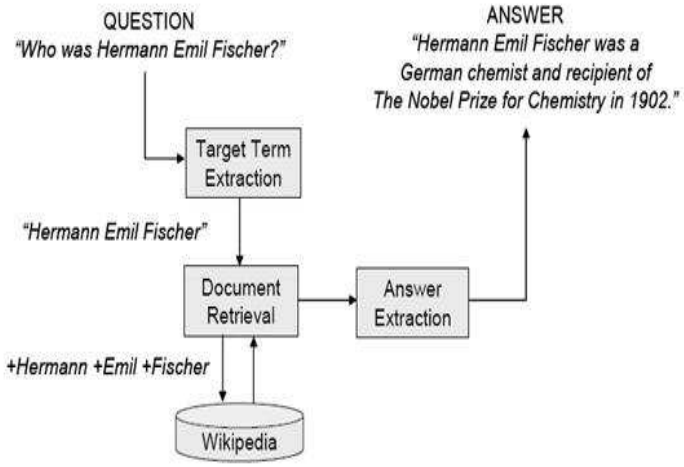

Fig. 2. Process for answer definition questions [7]

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 WAP technology in Question Answering

Wireless Application Protocol (WAP) is a secure specification that allows users to access information instantly via handheld wireless devices such as mobile phones, pagers, two-way radios, Smart phone and communicators.

WAP is designed to be user-friendly and innovative data applications for mobile phones easily. There are three types of terminals have been defined [12]:

- Feature phones, which offer high voice quality with the capability of text messaging and Internet browsing.
- Smart phones, with similar functionality but with larger display.
- The communicator, which is an advanced terminal designed with the mobile professional in mind, similar in size to a palm-top with a large display.

WAPs that use displays and access the Internet run what are called micro browsers; browsers with small file sizes that can accommodate the low memory constraints of handheld devices and the low-bandwidth constraints of a wireless-handheld network.

WAP uses Wireless Markup Language (WML), which includes the Handheld Device Markup Language (HDML) developed by Phone.com. WML can also trace its roots to eXtensible Markup Language (XML). A markup language is a way of adding information to your content that tells the device receiving the content and what to do with it. The best known markup language is Hypertext Markup Language (HTML). Unlike HTML, WML is considered a Meta language. Basically, this means that in addition to providing predefined tags, WML lets you design your own markup language components. 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. WML is scalable from two-line text displays up through graphic screens found on items such as smart phones and communicators.

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. Because WAP is fairly new, it is not a formal standard yet. It is still an initiative that was started by Unwired Planet, Motorola, Nokia, and Ericsson.

There are three main reasons why wireless Internet needs the Wireless Application Protocol:

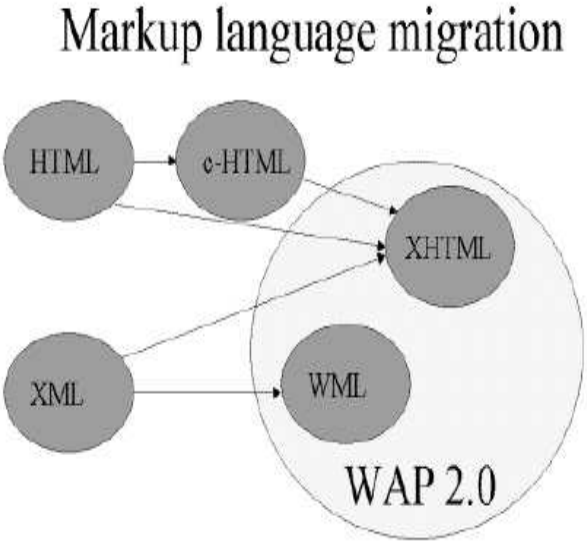

Fig. 3. Migration of Markup language

- Transfer speed: most cell phones and Web-enabled PDAs have data transfer rates of 14.4 Kbps or less. Compare this to a typical modem, a cable modem or a DSL connection. Most Web pages today are full of graphics that would take an unbearably long time to download at 14.4 Kbps. In order to minimize this problem, wireless Internet content is typically textbased in most cases.
- Size and readability: the relatively small size of the LCD on a cell phone or PDA presents another challenge. Most Web pages are designed for a resolution of 640x480 pixels, which is fine if you are reading on a desktop or a laptop. The page simply does not fit on a wireless device's display, which might be 150x150 pixels. Also, the majority of wireless devices use monochrome screens. Pages are harder to read when font and background colors become similar shades of gray.
- Navigation: navigation is another issue. You make your way through a Web page with points and clicks using a mouse; but if you are using a wireless device, you often use one hand to scroll keys.

WAP takes each of these limitations into account and provides a way to work with a typical wireless device.

Here's what happens when you access a Web site using a WAP-enabled device:

- You turn on the device and open the mini-browser.

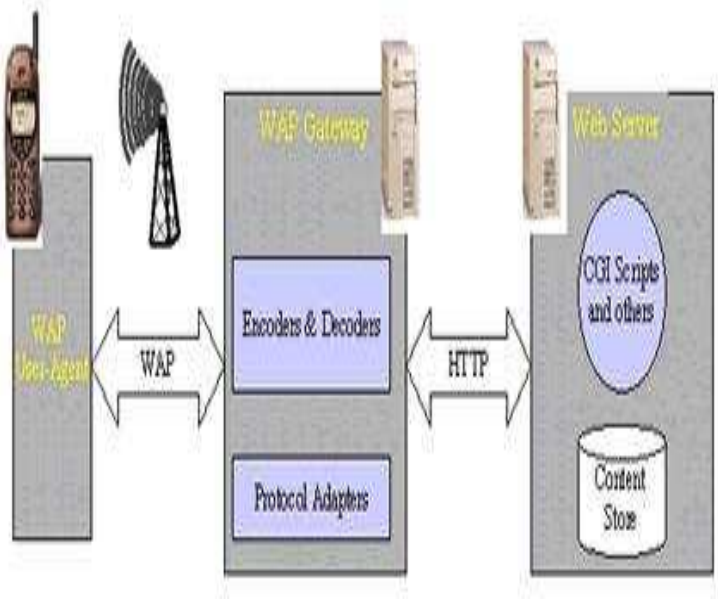

Fig. 4. WAP Technology Infrastructure

- The device sends out a radio signal, searching for service.
- A connection is made with your service provider.
- You select a Web site that you wish to view.
- A request is sent to a gateway server using WAP.
- The gateway server retrieves the information via HTTP from the Web site.
- The gateway server encodes the HTTP data as WML.
- The WML-encoded data is sent to your device.
- You see the wireless Internet version of the Web page you selected.

Although WML is well suited to most mundane content delivery tasks, it falls short of being useful for database integration or extremely dynamic content. PHP fills this gap quite nicely-integrating into most databases and other Web structures and languages. It's possible to "crossbreed" mime types in Apache to enable PHP to deliver WML content. WML pages are often called "decks". A deck contains a set of cards. A card element can contain text, markup, links, input-fields, tasks, images and more. Cards can be related to each other with links.

When a WML page is accessed from a mobile phone, all the cards in the page are downloaded from the WAP server. Navigation between the cards is done by the phone computer (inside the phone) without any extra access communications to the server.

5 Application mobile

As we mentioned at the begining, our proposal is the combination of mobile technologies and web technologies. First, we have development a mobile application (as you can see in figure 5) based on WAP technology. This application allows users to use at anytime and anyplace at very low cost, 2 cents per search. Furthermore, this application is available for most types of mobile phones. The figure 5 shows the main interface, as well as the request and response from the user's search.

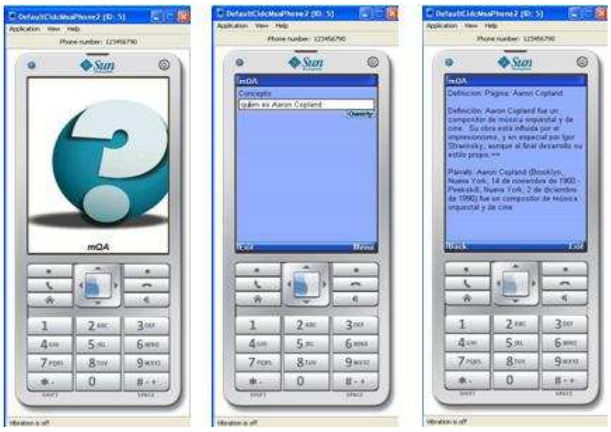

Fig. 5. Mobile application through WAP Technology

On the other hand, the figure 6 shows how our application mQAB can be accesed from web via iPhone through Wi-Fi. This is another channel of access to our application via wireless network. This feature allows our application covering all existing wireless and mobile devices.

6 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

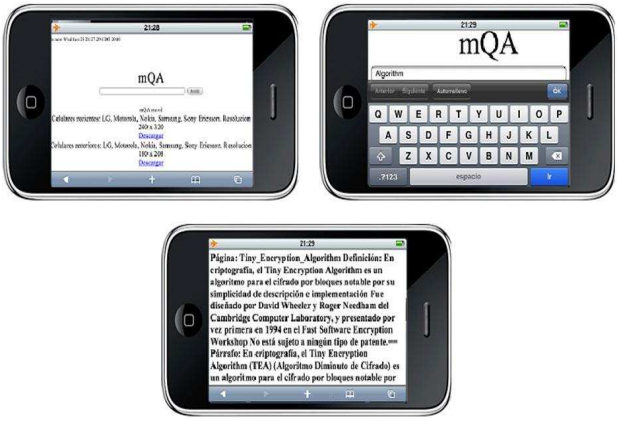

Fig. 6. Mobile application through iPhone

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).

7 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.

8 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/
12. J. AlSadi, B. AbuShawar, *MLearning: The Usage of WAP Technology in E-Learning*, International Journal of Interactive Mobile Technologies/Vol. 3, (2009)

Data Warehouse Development to Identify Regions with High Rates of Cancer Incidence in México through a Spatial Data Mining Clustering Task.

Joaquin Pérez Ortega[1],
María del Rocío Boone Rojas[1,2],
María Josefa Somodevilla García[2],
Mariam Viridiana Meléndez Hernández[2]

1 Centro Nacional de Investigación y Desarrollo Tecnológico, Cuernavaca Mor. Mex.
2 Benemèrita Universidad Autónoma Puebla, Fac. Cs. de la Computaciòn, México.
jperez@cenidet.edu.mx,{rboone,mariasg}@cs.buap.mx,*mvmh_099@hotmail.com*

Abstract

Data warehouses arise in many contexts, such as business, medicine and science, in which the availability of a repository of heterogeneous data sources, integrated and organized under a unified framework facilitates analysis and supports the decision making process. These data repositories increase their scope and application, when used for data mining tasks, which can extract useful knowledge, new and valuable from large amounts of data.

This paper presents the design and implementation of population-based data warehouses on the incidence of cancer in Mexico; based on the conceptual level multidimensional model and the ROLAP model *(Relational On-Line Analytical Processing)* at the implementation level.

A data warehouses is built, to be used as input for clustering data mining tasks, in particular, the k-means algorithm, in order to identify regions in Mexico, with high rates of cancer incidence.

The identified regions, as well as, the dimension related to the geographic location of the municipalities and their rate of incidence of cancer, are processed by IRIS, a Geographic Information System, developed at the National Institute of Statistics, Geography and Informatics of Mexico.

1 Introduction

Data warehouses arise in many contexts, such as business, medicine and science, in which the availability of a repository of heterogeneous data sources, integrated and organized under a unified framework facilitates analysis and supports the decision making process. These data repositories increase their scope and application, when used for data mining tasks, which can extract useful knowledge, new and valuable from large amounts of data.

Data warehouses have been applied mainly in the commercial and business areas [3] and more recently there have been some applications in the Health field

[16] [17] and the trend towards its integration with various technologies [11] [16].

Moreover, according to the literature, the use of data mining systems applied to the analysis of massive databases of health on a population basis has been limited, it is noteworthy work: *Constructing Over Dendrogram Matrix Detail view + Views.* [6], *Application of data mining techniques to databases population of cancer* [1], *Subgroup discovery in cervical cancer using data mining Techniques* [18] and *Data mining for cancer management in Egypt* [10]. In the case of Mexico, to the best of our knowledge, the work that has been developed at the *Centro Nacional de Investigación y Desarrollo Tecnológico* and BUAP, are the first ones in this field.

This work has been preceded by other works which has been done on the incidence of other cancers such as stomach and lung [15]. It is part of a larger project doomed to make proposals for improving the k-means algorithm in various aspects such as effectiveness and efficiency, reported in [12], [13] and [14] and its application in the Health field.

This article presents the data warehouse design and integration for the development of a data mining task on cancer incidence by regions in Mexico, based on the integration of complementary technologies such as clustering and geographical information systems. As a study case, the results for the incidence of cervical cancer are presented, which has been of special interest, since in Mexico, cervical cancer is the leading cause of cancer death in women [11].

The report is organized as follows, followed by this introduction, Section 2 presents the description of data sources and process design and implementation of data warehouse, Section 3 provides an overview of each application. In Section 4, results for the case of cervical cancer and its visualization by GIS INEGI IRIS [5] are included. Finally, in Section 5, conclusions and perspectives of this work are presented.

2 The Data Warehouse

The process of collecting and integrating data warehouse on cancer incidence by region in Mexico, required to select the data sources necessary to accomplish the task of data mining. This section describes the data sources and the conceptual design based on the multidimensional model and also, the implementation of the data warehouse under the ROLAP approach.

2.1 The Data Sources

In the study, the processed databases have been derived from official records of the National Institute of Public Health (INSP) and the National Institute of Statistics, Geography and Informatics (INEGI) of Mexico.

Data on cancer incidence were obtained through subsystem Remote Consultation System for Health Information (SCRIS) of the INSP [9]. In

particular, the databases were queried for cases of mortality cancer and results were configured by considering levels of aggregation such as: National States, division (Jurisdiction, Municipalities), year, age range, gender and causes (including tumors).

The information on the population and the actual geographical location of the municipalities was obtained from INEGI official databases, through its Geographic Information System IRIS, which has statistical information covering a wide geographical number of subjects, demographic, social and economic; also includes aspects of the physical environment, natural resources and infrastructure. This wealth of statistical and geographical data was obtained through various activities such as conducting population and housing census and economic census and the generation of basic cartography and census.

The information in the databases of the above institutions are integrated into a data warehouse (see Fig. 1), and according to the conventions in the area of health, for this study, only the municipalities with more than one hundred thousand inhabitants were considered.

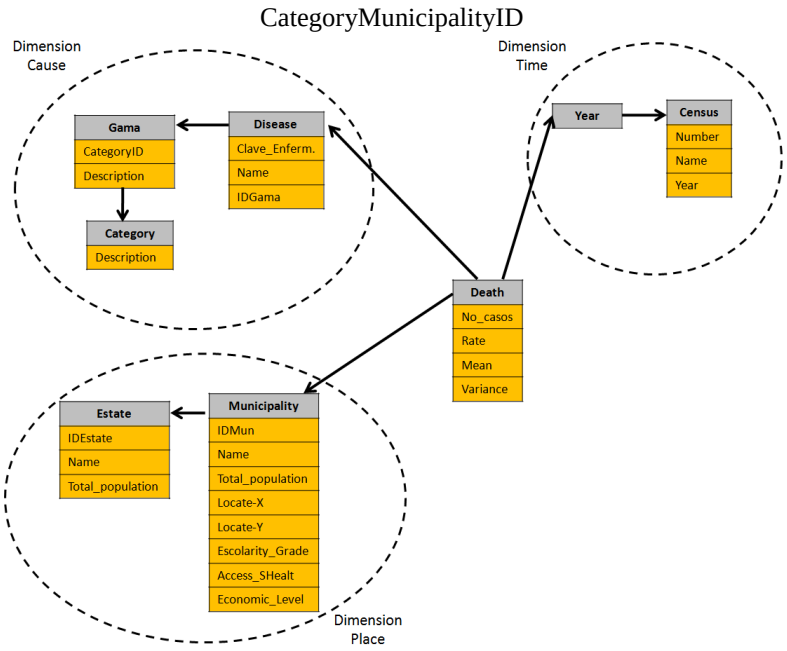

Fig. 1 Multidimensional Model Data Warehouse on the incidence of cancer in Mexico.

2.2 Data Warehouse Multidimensional Model for a population-based incidence of cancer in Mexico.

According to [4] the conceptual data model most widely used for data warehouses is the multidimensional model. The data are organized around the *facts* that have attributes or *measures* that may be more or less detail according to certain *dimensions*. In our case, the data warehouse design at the conceptual level is based on the multidimensional model, in which the dimensions can be distinguished as CAUSE, TIME, and PLACE. In this case, it is considered that a country has the basic fact, "deaths" that may have associated attributes such as number of cases, incidence rate, mean, variance, etc.. Fact can be detailed in several dimensions such as cause of death, place of death, date of death, etc. In Fig. 1 shows the facts "deaths" and three dimensions with various levels of aggregation. The arrows can be read as "is added". As shown in Fig. 1, each dimension has a hierarchical structure but not necessarily linear. When the number of dimensions cannot exceed three represent each combination of levels of aggregation as a cube.

The cube is made up of boxes with one box for each possible value from each dimension to the corresponding level of aggregation. On this "view", each box represents a fact. Fig. 2 shows a three dimensional cube corresponding to the fact: "According to the 2000 census, the town of Atlixco, there were 15 deaths from cervical cancer" in which the dimensions Cause, Place and Time have been added by type of disease (cancer), Municipality and Census. The representation of a fact corresponds therefore to a square in the cube. The value of the box is the observed (in this case is the number of deaths).

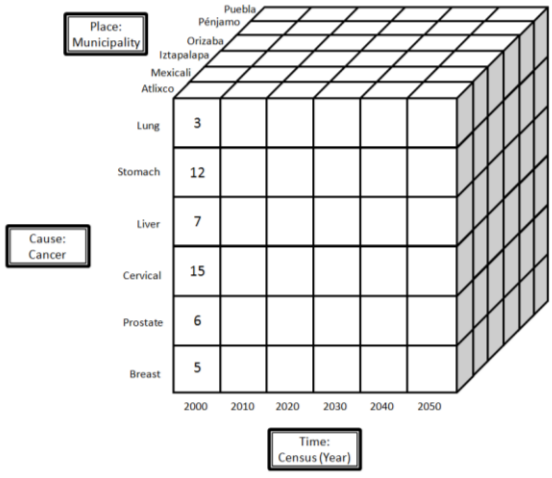

Fig. 2 Display of a fact in a multidimensional model

2.3 Data warehouse scheme ROLAP (Relational OLAP) implementation of population-based cancer incidence in Mexico.

One of the most efficient ways to implement a multidimensional model using relational databases is based on the ROLAP model [4]. In our case, the tables for the ROLAP model have the following schemes:

Snowflake Tables

Dimension Cause

DISEASE (Clave_Enfermedad, name, IdGama, CategoryID)
GAMA (IdGama, CategoryID, Description)
CATEGORY (CategoryID, Description)

Place dimension

STATE (Clave_Estado, name, población_total)
MUNICIPALITY (Clave_Municipio, Clave_Estado, name, población_total, Loc_x, Loc_y, extension, tipo_zona, nivel_socioeconómico)

Time dimension

YEAR (Idan)
CENSUS (IdCenso, Idan, number, name)

Fact Tables

DEATH (IdEnfermedad, IdCenso, IdMunicipio, no_casos, rate, mean, variance)

Star Tables

TIME (Idan, IdCenso)
CAUSE (IdEnfermedad, IdGama, CategoryID)
PLACE (IdCiudad, IdMunicipio)

3 Data Mining Application on Cancer Incidence

The implemented data warehouse has been used to develop a data mining task space based on the integration of additional technologies to the data warehouse, such as clustering and Geographic Information Systems, which in this case are very suitable, to identify and display areas with incidence of cancer in Mexico. The following provides a general description of the integration process of technologies and tools (Fig. 3) made for this application.

The data warehouse integrates the following information for our application: the component space that allows viewing of the regions of municipalities, population data such as the death rate and incidence rate and the time component, which in this case is the census year.

The IRIS GIS INEGI [5], through your options allows the recovery of population data and the real location of the municipalities, which are integrated into the data warehouse.

Since IRIS stores geographical representation of municipalities in the vector format standardized *"shape"* and by means of polygons, there is the need for a process of transfer of forms and formats in order to have a numerical representation of each municipality, in this case, corresponds to a point on the municipality center location, which is accomplished primarily through the tools of ESRI's ArcInfo GIS.

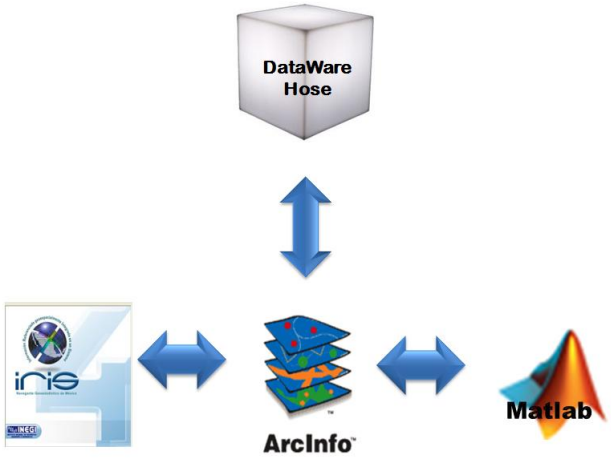

Fig. 3 Integration of Technology and Data Mining Tools

Given the numerical representation of each municipality through a point (x, y), along with its rate of incidence of cancer, the Matlab programming environment and its implementation of *k-means* algorithm [2] [7] is used *to* generate patterns / groups of municipalities and the corresponding centroids.

Once you have the above results, it is again necessary to transfer digital data format to format *shape*, a process similar to above using ArcInfo tools, allowing viewing through GIS IRIS.

Finally, the groups of municipalities and their corresponding centroids, are passed as GIS layers to IRIS, for display on the geographic map of Mexico.

4 Results and visualization with IRIS

In this project we have done grouping tasks according to the affinity of location and incidence rate of the municipalities. Series of experimental tests on the data stores in cities with more than 100.000 inhabitants were carried out. Size groups were considered k = 5, 10, 15, 20 and 30. The best result was obtained for k = 20.

As a case study, this paper presents the results obtained by k-means algorithm in Matlab for the cervical cancer data warehouse. Fig. 4 provides the visualization of the 20 regions identified.

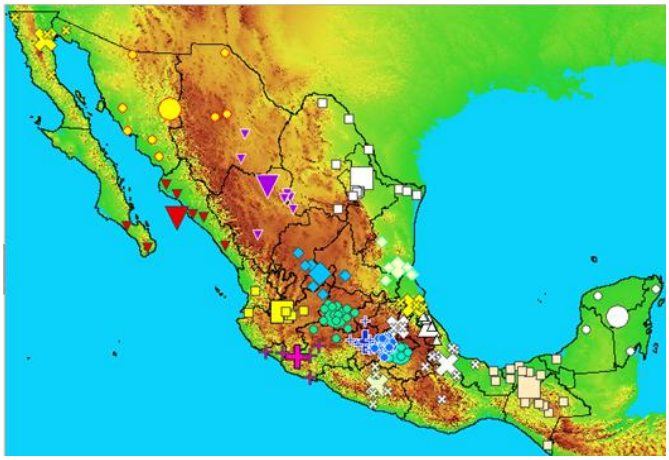

Fig. 4 Regions of the Municipalities with an incidence of Cervical Cancer.

From the results, we distinguish the groups spearheading the three municipalities with higher incidence rates: *Atlixco*, *Apatzingán* and *Tapachula (Chiapas)*. In Fig. 5 the detail of the display of the group corresponding to the region of Chiapas and the incidence of cervical cancer is shown. Table 1 provides data for the previous group, and statistical measures for the mean and standard deviation.

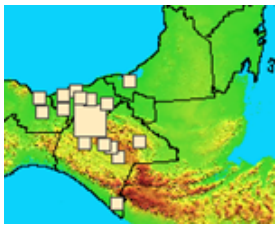

Fig. 5 Tapachula Chiapas Group

The groups identified with high incidence rates: *Tapachula* and *Apatzingan* match municipalities identified in other studies [4] and correspond to the population characteristics, identified in the work of the medical field [8], [15] such situations such as poverty, lack of preparation and access to effective health services and the initiation of sexual activity at an early age. This allows us to

assert that the grouping is made valid. On the other hand, the study allowed discovering other municipalities that had not been identified in other research, such as the group of Atlixco, in particular showing the highest incidence rate in the country (see table 2).

Table 1 Municipalities Incidence Rates of Cervical-Uterine Cancer

State	Municipality	Population	Deaths	Rate
Chiapas	Tapachula	271674	27	9.93
Veracruz-Llave	Coatzacoalcos	267212	23	8.60
Veracruz-Llave	Minatitlán	153001	13	8.49
Chiapas	Comitán de Domínguez	105210	8	7.60
Chiapas	San Cristóbal de las Casas	132421	9	6.79
Tabasco	Comalcalco	164637	11	6.68
Tabasco	Cárdenas	217261	11	5.06
Tabasco	Huimanguillo	158573	8	5.04
Chiapas	Tuxtla Gutiérrez	434143	21	4.83
Tabasco	Cunduacán	104360	5	4.79
Campeche	Carmen	172076	8	4.64
Tabasco	Macuspana	133985	6	4.47
Tabasco	Centro	520308	23	4.42
Chiapas	Ocosingo	146696	2	1.36
Average				5.91
Standard deviation				2.23

In order to perform a global analysis of our results, Table 2 provides information of the ten municipalities with the highest incidence rate in the country.

Table 2 Top Ten Municipalities Incidence Rates of Cervical-Uterine Cancer

Key	State	Municipality	Population	Deaths	Rate
21019	Puebla	Atlixco	117111	15	12,80
16006	Michoacán	Apatzingán	117949	13	11,02
07089	Chiapas	Tapachula	271674	27	9,93
17006	Morelos	Cuautla	153329	14	9,13
28021	Tamaulipas	El Mante	112602	10	8,88
06007	Colima	Manzanillo	125143	11	8,78
30039	Veracruz-Llave	Coatzacoalcos	267212	23	8,60
18017	Nayarit	Tepic	305176	26	8,51
30108	Veracruz-Llave	Minatitlán	153001	13	8,49
30118	Veracruz-Llave	Orizaba	118593	10	8,43
General Mean					4.70
Standard Deviation					1.95

Figure 6, illustrates the location of previous incidence rates compared to the national average and the corresponding standard deviation.

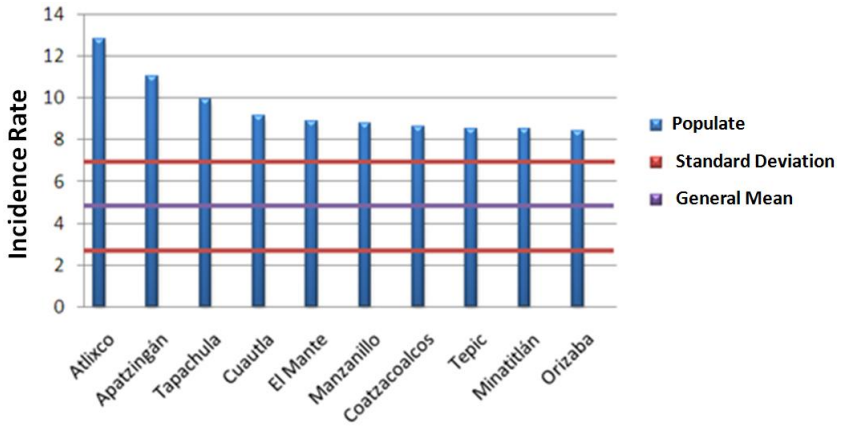

Figure 6. Top Ten municipalities incidence rates.

5 Conclusions

Multidimensional model for conceptual design of the data warehouse, turned out to be very appropriate, since this model is easily scalable and allows analysis of the information under different perspectives. It is expected that future studies process other variables, related to the municipalities, included in this design, such as socioeconomic status, type of region, gender and access to health services, among others. Moreover, the implementation of data warehouse based on the ROLAP model has allowed taking advantage of the facilities developed for relational databases. In addition, it is expected that the design and implementation carried out in the data warehouse can be used in other applications.

The processing of the spatial component of our data warehouse, using the IRIS GIS INEGI, has resulted in a high quality visual representation of our results, based on the actual physical location of the municipalities and on a map of the topography of the Republic Mexican INEGI. Also experience and learning has been gained on transfer of shapes (polygons, points) techniques and formats *(Number-shape)* through ArcView GIS tools.

Currently we are working to complete studies in other cancer types. Besides, data mining tasks will be developed on the incidence of conditions such as diabetes, influenza and cardiovascular diseases, among others.

Acknowledgement. R. Boone expresses her gratitude to Ms. Rocío Pérez Osorno from INEGI, Puebla. (Graduated from the Faculty of Cs. Computing, BUAP) for advice and support in plotting the results of this work through the IRIS GIS.

References

1. Barrón Vivanco M. Arandine, Pérez O. J., Miranda H. Fátima, Pazos R., XII Congreso de Investigación en Salud Pública, *Aplicación de técnicas de minería de datos a bases de datos poblacionales de cáncer*, CENIDET, México, Secretaría de Saúde do Estado de Pernambuco, Brasil, Abril (2007).
2. Forgy E. "Cluster analysis of multivariate data: Efficiency vs. Interpretability of classification", Biometrics, vol. 21, pp.768-780.1965
3. Hernández-Orallo J., Ramiréz-Quintana M. J., Ferri-Ramiréz C., Introducción a la Minería de Datos, Ed. Pearson Prentice Hall, Madrid (2004).
4. Hidalgo-Martínez Ana C. El cáncer cérvico-uterino su impacto en México. Porqué no funciona el programa nacional de detección oportuna. Revista Biomédica, Centro Nal. De Investigaciones Regionales Dr. Hineyo Noguchi, UADY, 2006, México.
5. IRIS 4. http://mapserver.inegi.gob.mx. SNIEG Sistema Nacional de Información Estadística y Geográfica.
6. Jin Chen, MacEachren, Alan M., Peuquet, Donna. Constructing Overview+Detail Dendogram Matrix Views. IEEE Transactions on Visualization & Computer Graphics., Vol. 15, Issue 6, p889-896, Dec. 2009.
7. 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.
8. Martínez M. Francisco Javier. Epidemiología del cáncer del cuello uterino. Medicina Universitaria 2004, 39-46. Vol. 6, N. 22, UANL, México.
9. NAIIS Instituto Nacional de Salud Pública, SCRIS, Mortalidad, http://sigsalud.insp.mx/naais/, Cuernavaca, Morelos, México, (2003).
10. Nevine M. Labib, Michael N. Malek: Data Mining for Cancer Management in Egypt. Transactions on Engineering, Computing and Technology V8 October 2005: (ISSN 1305-5313).
11. Pérez-C. Nelson, Abril-Frade D.O. Estado Actual de las Tecnologías de Bodegas de Datos Espaciales. Ing. E Investigación. Vol.27, No. 1, Univ. Nal. De Colombia. 2007.
12. Pérez-O. J.,1, R. Pazos R, L. Cruz R.,G. Reyes S. "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.
13. Pérez-O. J.2, 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 cancer. 2do Taller Latino Iberoamericano de Investigación de Operaciones, Mèxico, 2007.
14. Pérez-O. J.3, Rocío Boone Rojas, María J. Somodevilla García. Research issues on K-means Algorithm: An Experimental Trial Using Matlab., Advances on Semantic Web and New Technologies". Vol 534. http://ceur-ws.org/.
15. Rangel-Gómez, G. Lazcano-Ponce,E. Palacio-Mejía, Cáncer cervical, una enfermedad de la pobreza: diferencias en la mortalidad por áreas urbanas y rurales en México, http:// www.insp.mx/salud/index.html.
16. Scotch,Matthew, Parmato B. Monaco, V. Evaluation of SOVAT: An OLAP-GIS decision support system for community health assessment data analysis. BMC Medical Informatics & Decisión Making Vol. 8 (1-12). 2008.
17. Simonet, A., Landais, P. Guillon D.A multi-source Information System for end-stage renaldisease. Comptes Residus Biologies, 2002, Vol. 325 I4., p515.
18. Thangavel K. Jaganathan P. and Esmy P. O., *Subgroup Discovery in Cervical Cancer Analysis Using Data Mining Techniques*, Departament of Computer Science, Periyar University: Departament of Computer Science and Applications, Gandhigram Rural Institute-Deemed University, Gandhigram: Radiation Oncologist , Christian Fellowship Community Health Centre, Tamil Nadu, India: AIML journal, Vol(6), Issue(1), January, 2006.

An Approach of Crawlers for Semantic Web Application

José Manuel Pérez Ramírez[1] , Luis Enrique Colmenares Guillen[1]

[1] Benémerita Universidad Autónoma de Puebla,
Facultad de Ciencias de la Computación,
BUAP – FCC, Ciudad Universitaria,
Apartado Postal J-32,
Puebla, Pue. México.
{ mankod, lecolme}@gmail.com

Abstract. This paper presents a proposal for a system capable of retrieval information from the processes generated by the system Yacy. The information retrieved will be used in the generation of a knowledge base. This knowledge base may be used in the generation of semantic web applications.

Keywords: Semantic Web, Crawler, Corpora, Knowledgebase.

1 Introduction

A knowledgebase is a special type of database for managing knowledge. It provides the means to collect organize and recover knowledge in a computed way. In general, a knowledgebase is not a static set of information it is a dynamic resource that maybe have the ability to learn. In the future, Internet will be a complete and complex knowledgebase, already known as *semantic web* [1].

Some examples of knowledge base are: a public library, an information database related to a specific subject, *Whatis.com, Wikipedia.org, Google.com, Bing.com and Recaptcha.net.*

Investigate related to ***Generation Automatic of a specialized corpus*** from the Web is present in [2], this investigate have a reviews of methods to process knowledgebase that generates specialized *corpus.*

In section 2 we present related work to *semantic web* in order to comprehend the benefits that may be obtained by elaborating them.

In Section 3 we describe the challenges and we explain the problems that could be have if you tried to use Google Search for getting information or tried to retrieval information of queries to Google.

Section 4 the methodology to use for solving the problem. And section 5, conclusions and ongoing work.

We continue this paper present a form abstract to describe a **Query Processing on the Semantic Web** [8] is as follows Fig. 1

1. A query with a data type.
2. A server that sends queries to the servers decentralized indexing. The content found on the servers is similar to indexing a book index indicates which pages contain the words that match the query.
3. The query travels to the servers where documents stored documents are retrieved are generated to describe each search result.
4. The user receives the results of its semantic search which has already been processed in the semantic web server.

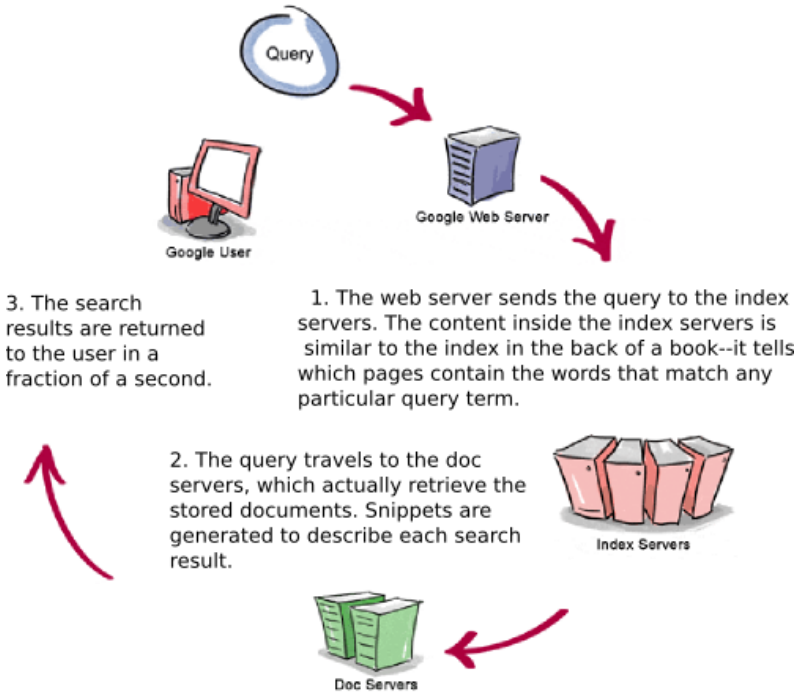

Fig. 1. Querying the Semantic Web.

2 Related Work

Nowadays, the investigation related to ***retrieval information*** on the web has a different result like: knowledgebase, web sites dedicated to retrieval information, Wikipedia, Twine, Evri, Google, Vivísimo, Clusty, etc.

An example of a company that working with "retrieval information" is Google Inc, one of their products is Google Search this *web search engine* is the one of the most-used search engine on the Web [9], Google receives several hundred million queries each day through its various services [10].

This kind of example it's necessary for the following analogy: *For what reason Google doesn't put their information of their knowledgebase under domain public?*
And the answer it's very simple: because their *information* or their *knowledgebase it's* ***money.***

In section 3 we explain some form of extract information of Google Search only a protected few of information it's impossible retrieval many information of Google Search whit the idea to generate knowledgebase this because Google protects their information of their queries.

Another kind of knowledgebase are:

2.1 Wikipedia

A specific case is Wikipedia, a project to write a free communitarian encyclopedia in all languages. This project have 514 621 articles today. The quantity and quality of the articles present an excellent knowledgebase for the creation of semantic webs.
We present some ways to obtain semantic information from Wikipedia: from its structure, from the collected notes of the people that contributes and from the existent links in the entries.

2.2 Twine

Twine is a tool for storage, organizes and shares information, all of it with an intelligence provided by the platform that analyzes the semantic of the information and classifies automatically [7]. The main idea is to save users from labeling and connecting related content and leave this work to Twine, bringing more value and storage the contents next to the information about its meaning.

3 Challenges

The principal challenge is development a system with the capacity of works with Yacy for retrieval information of Indexing Process and generate information this information will be essential for produce knowledgebase.

We present in the figure 5 all modules of yacy, so the module to development will be works with some of these modules.

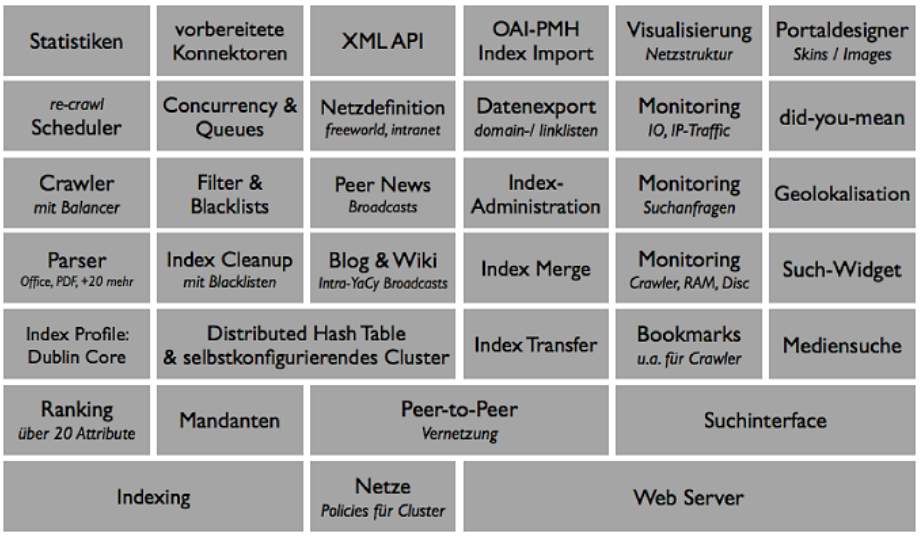

Figure 5. Components of Yacy

The principal question is:

What we can do to get information under domain public.

It's very simple we use the very popular Wikipedia

Wikipedia is a project of the Wikimedia Foundation. More than 13.7 million of its articles have been drafted in conjunction with volunteers from all over the world and practically every one of them may be edited by any person that may have access to Wikipedia. Actually it is the most popular reference work on the internet.

This project of dynamic content like Wikipedia illustrates the information that have great potential to be exploited.

Otherwise Google Search is one of the most-used search engine provides at least 22 special features beyond the original word-search capability. These include synonyms, weather forecasts, time zones, stock quotes, maps, earthquake data, movie showtimes, airports, home listings, and sports scores.

And maybe you could be thinking:

*For what reason the people don't use a Google Search for get all the **knowledgebase** about topic specific and this **knowledgebase** could be export to file of text plan with the possibilities of management this and generate corpus.*

Very simple is the answer because the information of Google is their information and gold for company.

It the past Google Inc. allowed the retrieval information from any kind of query[3].

Google allowed the retrieval information based on their form and methods like ***University Research Program for Google Search*** [10] but any kind of answered we get of this project when we make the inscription to this program.

Another way to exploit Google Search knowledge is using scripts, APIS [3], programming languages such as AWK, development tools like SED or GREP, all of them analyzed in [2] but with few results and we need a lot of information for create *knowledgebase*.

3.1 Considerations

1. Create a module with the goal to connect this with YACY and retrieval information of their crawlers.
2. Export a set of information related with a topic in plain text.
3. Management information of web site like Wikipedia.org.
4. Index the content of this kind of retrieval information in storage local.
5. Public the module in the web and share the *knowledgebase.*

4 Methodology

This section gives a description of the project taking into consideration the design that will be used to give a solution to the problem of creating the module.

4.1 Project description

The obtained results of the module that connected with Yacy will be used to create semantic webs, corpus and any other project that needs information in a plain text about web content.

Described below are a series of procedures to follow that use as a methodology to implement within the project.

A) Check the modules of Yacy

B) Check the logistic and architecture of Yacy

C) Check the form that Yacy create their crawlers

D) Think in a form of create the Module capable of manage the information of the crawler and generate *knowledgebase*

E) Some of the polices described above are implemented in YaCy [6], the variant to use is the implementation of the JXTA[5] tool and the URI and RDF policies that allow to structure and outline the results, to finally present then in a *semantic* way or *knowledgebase*.

4.2 Development platform

This work is done with YaCY, which is a free distribution search engine, based on the principles of the peer to peer (P2P). Its core is a program written in Java that it's distributed in hundreds of computers, from September 2006. It's called YaCy-peer. Each YaCy-peer is an independent crawler that navigates trough the Internet, and analyzes and indexes web pages found. To storages the indexation results in a common database (called index) which is shared with other YaCy-peers using the principles of the P2P networks [4].

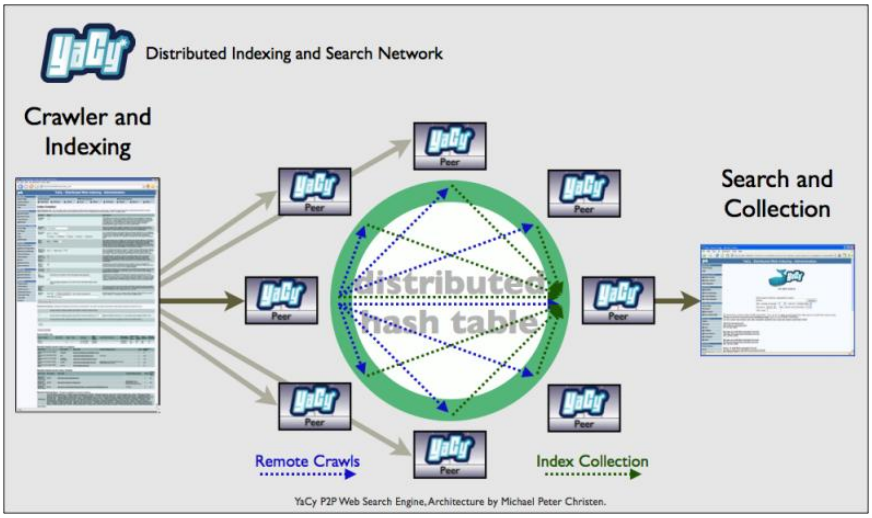

Fig. 2 Distributed indexing process

Compared to semi-distributed search engines, the YaCy-network has a decentralized architecture. All of the YaCy-peers are equal and there is no central server. It may be executed in Crawling mode or as a local proxy server. The figure 2 shows a diagram that describes the distributed process of indexation and the search in the network for the YaCy crawler.

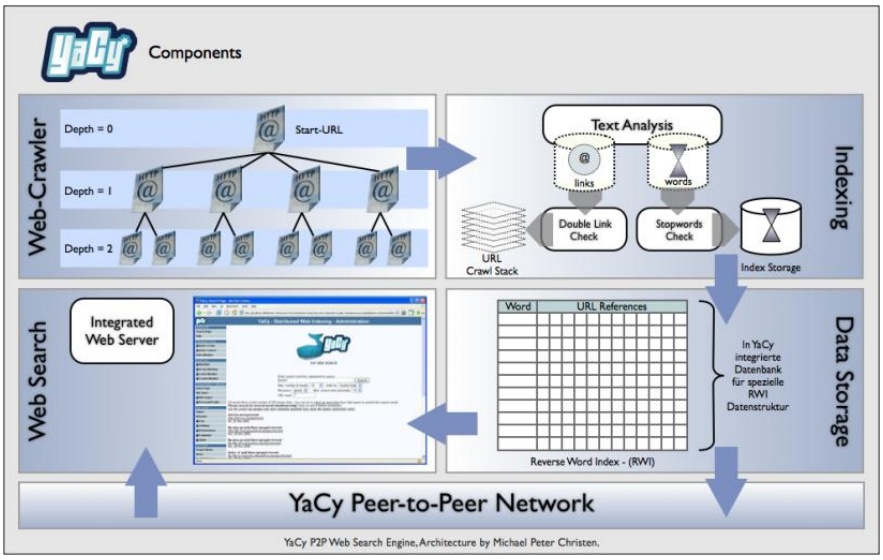

Fig. 3. Distributed indexing process

The figure 3, to have the main components of YaCy, and the process that exists among the web search, web crawler, the indexing and data storage processes.

5 Conclusions and ongoing work

In this section present some the conclusions and results that are expected of project and the future work.

1. Index all content of Wikipedia.
2. Storage this content.
3. Present the content of Wikipedia by topic in a web site.
4. Use a tagged of text for share the information with tags.
5. Present the module and their code on a web site
6. Share *knowledgebase* extract of Wikipedia

References

1. Definition of knowledgebase http://searchcrm.techtarget.com/definition/knowledge-base
2. Alarcón, R., Sierra, G., Bach, C. (2007). "Developing a Definitional Knowledge Extraction System". En Vetulani, Z. (ed.), Actas del 3er *Language & Technology Conference. Human Language Technologies as a Challenge for Computer Science and Linguistics*. Poznan, Universidad Adam Mickiewicza: pp. 374-378.
3. Google Hacks, Second Edition, 2004, O'Reilly Media.
4. S. Rhea, B. Godfrey, B. Karp, J. Kubiatowicz, S. Ratnasamy, S. Shenker, I. Stoica, and H. Yu. OpenDHT: a Public DHT Service and its Uses. SIGCOMM' 05, Philadelphia, Pennsylvania, USA, august 21-26, (2005).
5. http://www.jxta.org (2010).
6. http://yacy.net/ (2010).
7. http://www.twine.com/ (2010).
8. Query Processing on the Semantic Web Heiner Stuckenschmidt, Vrije Universiteit Amsterdam
9. http://www.alexa.com/siteinfo/google.com+yahoo.com+altavista.com (2009)
10. http://searchenginewatch.com/showPage.html?page=3630718 (2008)
11. http://research.google.com/university/search/ (2010)

Decryption Through the Likelihood of Frequency of Letters

Barbara Sánchez Rinza, Fernando Zacarias Flores, Luna Pérez Mauricio, and Martínez Cortés Marco Antonio

Benemérita Universidad Autónoma de Puebla,
Computer Science
14 Sur y Av. San Claudio, Puebla, Pue.
72000 México
brinza@cs.buap.mx, fzflores@yahoo.com.mx

Abstract. The method to decrypt the information using probability leads to a more thorough job, because you have to know the percentage of each of the letters of the language that is being analyzed here is Spanish. You can consider not only the probabilities of the letters also syllables, set of three, four letters and even words. Then you have this thing to do is make comparisons of the frequencies of cipher text and the frequencies of the language to begin to replace by a correspondence. And finally passing a scanner and find the decrypted text.

Keywords Probability, Decrypt.

1 Introduction

Cryptography is the science that alters the linguistic representations of a message [1]. For this there are different methods, where the most common is encryption. This science masking the original references of the information by a conversion method governed by an algorithm that allows the reverse or decryption of information. Use of this or other techniques, allowing for an exchange of messages that can only be read by the intended beneficiaries as 'consistent'. A consistent recipient is the person to whom the message is directed with the intention of the sender. Thus, the recipient knows the discrete coherent used for masking the message. So either have the means to bring the message to the reverse process cryptographic, or can infer the process that becomes a message to the public. The original information to be protected is called plaintext or cleartext. Encryption is the process of converting plain text into unreadable gibberish called ciphertext or cryptogram. In general, the concrete implementation of the encryption algorithm (also called figure) is based on the existence of key secret information that fits the encryption algorithm for each different use [2].

Decryption is the reverse process to recover the plaintext from the ciphertext and key. Cryptographic protocol specifies the details of how to use algorithms and keys (and other primitive operations) to achieve the desired effect. The set

of protocols, encryption algorithms, key management processes and actions of the users, which together constitute a cryptosystem, which is what the end user works and interacts. In this work, we must first have a ciphertext which must meet certain requirements, such a text should be bijective so that each element of the domain carries a single element of the condominium. In addition we must also take account of the rules of Kerckhoff [3].

2 Development work

2.1 Frequencies in Spanish

Is required to decrypt text using the odds as to how often they used certain letters in the alphabet, for this work only considered the Spanish language [5].

The frequencies of Spanish, which were used for this study were:

1. *Frequency triglyphs*
2. *Frequency of digraphs*
3. *Most common words*
4. *Frequency of letters at the beginning of words*
5. *Frequency of letters in Spanish*
6. *Frequency Words*

2.2 Triglyphs Frequencies

The letter frequency statistics may vary from one to another depending on the corpus author has chosen to develop them. Usually differences when the corpus is literary or consists of texts of different origins. Table 1 shows the frequency of each of the Spanish alphabet with their respective percentage.

High frequency letters		Medium frequency letters		Low frequency letters		Frequencies 0.5%
letter	freq.%	letter	freq.%	letter	freq.%	G, F, V, W
E	16,78	R	4,94	Y	1,54	
A	11,96	U	4,80	Q	1,53	
O	8,69	I	4,15	B	0,92	
L	8,37	T	3,31	H	0,89	
S	7,88	C	2,92			J, Z, X, K, N
N	7,01	P	2,76			
D	6,87	M	2,12			

Table 1. Frequency triglyphs

2.3 Most Frequent words

The vowels make up about 46.38% of the text. The high frequency letters account for 67.56% of the text. Mid-frequency points accounting for 25% of the text [4]. In the dictionary the most common vowel is A, but in written texts is the E because of prepositions, conjunctions, verbs, etc. The most common consonants are L, S, N, D, with about 30%. The less frequent six letters: V, N, J, Z, X and K (just over 1%). The average frequency of a Spanish word is 5.9 letters. The coincidence index for Spanish is 0.0775. In addition to solving the encryption table 2 we mentioned that we most frequently used words in a text of 10 000 words.

Most common words		Two-letter words		Three-letter words
Word	Frequency	Frequency	Word	Frequency
DE	778	778	QUE	289
LA	460	460	LOS	196
El	339	339	DEL	156
EN	302	302	LAS	114
QUE	289	119	POR	110
Y	226	98	CON	82
A	213	74	UNA	78
LOS	196	64	MAS	36
DEL	156	63	SUS	27
SE	119	47	HAN	19
LAS	114			

Table 2. Most frequent words of one, two and three letter

Next, table 3 shows the frequencies of the 4-letter words.

2.4 Frequency digraphs

The size of the corpus is 60,115 letters. The frequencies are absolute. The digraphs are read by row and column in that order. Below in table 4 shows the union digraphs are letters from letters.

2.5 Most common initial letter

The most frequent letters in Spanish that start a word are listed in Table 5

3 Results

The ciphertext is used as said it had to be bijective and have Kerckhoff rules and the decrypted text shown in Figure 1.

Four-letter words		Distribution of letters in literary texts			
Word	Frequency	E - 16,78%	R - 4,94%	Y - 1,54%	J - 0,30%
PARA	67	A - 11,96%	U - 4,80%	Q - 1,53%	
COMO	36	O - 8,69%	I - 4,15%	B - 0,92%	
AYER	25	L - 8,37%	T - 3,31%	H - 0,89%	
ESTE	23	S - 7,88%	C - 2,92%	G - 0,73%	
PERO	18	N - 7,01%	P - 2,77%	F - 0,52%	
ESTA	17	D - 6,87%	M - 2,12%	V - 0,39%	
AOS	14				
TODO	11				
SIDO	11				
SOLO	10				

Table 3. Frequency with four letters

4 Conclusions

We conclude that this method of decryption is good however would have to tweak a little more due to it depends on the text we have and how much text to decrypt was also observed that only decrypts an encrypted bijective. In this work, as seen in the results of Figure 1, which apply various processes, first see the probability of the lyrics in Spanish that are more frequent, then seen with the syllables that are more frequent in Spanish, and then with the last word and you miss the information, text analyzer, as shown in Figure 1 a large percentage of the information is decoded, but as mentioned in the top, this will depend have that much information to process it.

References

1. Liddell and Scott's Greek-English Lexicon. Oxford University Press. (1984)
2. Anaya Multimedia, Codigos Y Claves Secretas: Programas En Basic, Basado A Su Vez En Un Estudio Lexicogrfico Del Diario "El Pas", Mexico 1986.
3. Friedman, William F. And Callimahos, Lambros D., Military Cryptanalytics, Cryptographic Series, 1962
4. Part I - Volume 2, Aegean Park Press, Laguna Hills, Ca, 1985
5. Barker, Wayne G., Cryptograms In Spanish, Aegean Park Press, Laguna Hills, Ca.,

	A	B	C	D	E	F	G	H	I	J	K	L	M
A	12	14	54	64	15	5	8	4	10	8		41	30
B	11				5				14	1		12	
C	39		5		17			8	80			3	
D	32		1	2	84			1	30				
E	20	5	47	26	17	8	21	6	9	3		44	26
F	2				9				12			1	
G	12				12				5			1	
H	15				3				5				
I	43	8	42	29	40	5	8			1		14	16
J	4				5								
K					1								
L	44		5	5	35	1	3		28			9	5
M	32	10			42				30				
N	41	2	33	37	41	10	6	2	28	1		5	4
O	19	17	28	26	16	6	5	5	4	1		22	33
P	30		1		16				5			8	
Q													
R	74	1	12	10	94	1	12		45	1	1	6	15
S	32	2	18	15	57	3	2	4	41	1		5	7
T	60		1		67				35				
U	13	6	11	5	52	1	3		9			9	6
V	12			1	15				15				
W	1				1								
X			1		4								
Y	5	1	3	2	5	1	1					1	1

Table 4. Frequency of digraphs

letter	P	C	D	E	S	A	L	R	M	N	T			
frequency	1.1128	1.081	1.012	989	789	761	435	425	403	346	298			
letter	Q	I	H	U	G	V	F	O	B	J	Y	W	Z	K
frequency	286	281	230	219	206	183	177	169	124	47	27	19	2	1

Table 5. Frequency of initial letters

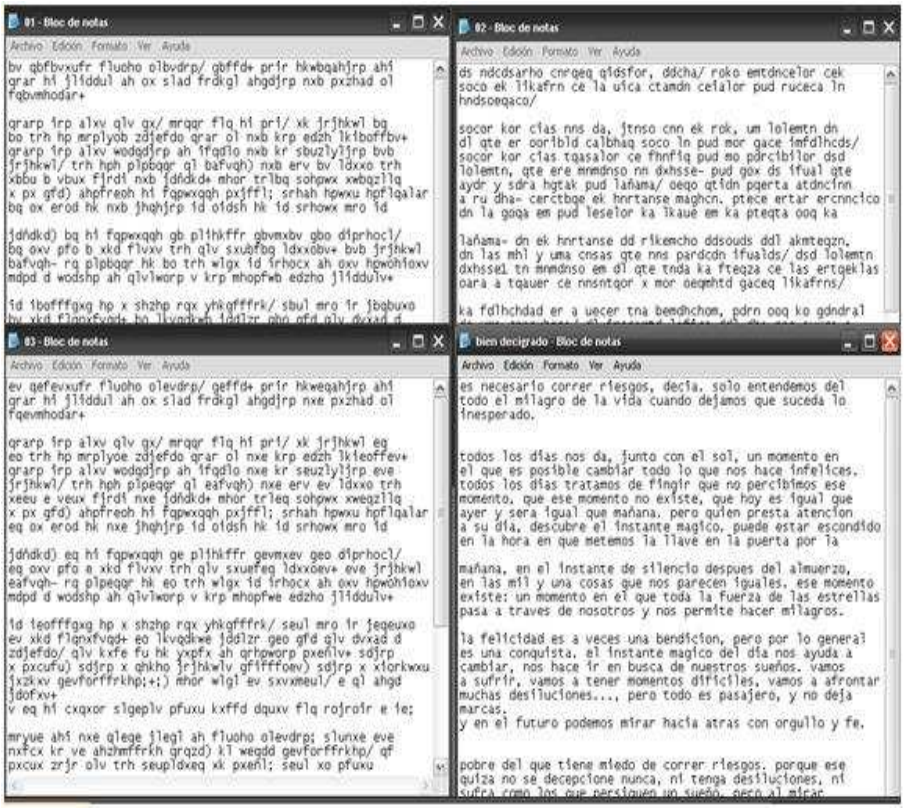

Fig. 1. with each of the texts worked, 01 encrypted text, 02 text one pass, 03 second pass the text, either original text decrypted