1st Symposium on Information Management and Big Data

8th, 9th and 10th September 2014
Cusco - Peru

PROCEEDINGS

Proceeding Editors: J. A. Lossio-Ventura and H. Alatrista-Salas

TABLE OF CONTENTS

The SIMBig 2014 Organization Committee confirms that full and concise papers accepted for this publication:

- Meet the definition of research in relation to creativity, originality, and increasing humanity's stock of knowledge;
- Are selected on the basis of a peer review process that is independent, qualified expert review;
- Are published and presented at a conference having national and international significance as evidenced by registrations and participation; and
- Are made available widely through the Conference web site.

Disclaimer: The SIMBig 2014 Organization Committee accepts no responsibility for omissions and errors.

SIMBig 2014 Organization Committee

GENERAL CHAIR

- Juan Antonio LOSSIO VENTURA, Montpellier 2 University, LIRMM, Montpellier, France
- Hugo ALATRISTA SALAS, UMR TETIS, Irstea, France

LOCAL CHAIR

- Cristhian GANVINI VALCARCEL Andina University of Cusco, Peru
- Armando FERMIN PEREZ, National University of San Marcos, Peru
- Cesar A. BELTRAN CASTAÑON, GRPIAA, Pontifical Catholic University of Peru
- Marco RIVAS PEÑA, National University of San Marcos, Peru

SIMBig 2014 Conference Reviewers

- Salah Ait-Mokhtar, *Xerox Research Centre Europa, FRANCE*
- Jérôme Azé, *LIRMM - University of Montpellier 2, FRANCE*
- Cesar A. Beltrán Castañón, *GRPIAA - Pontifical Catholic University of Peru, PERU*
- Sandra Bringay, *LIRMM - University of Montpellier 3, FRANCE*
- Oscar Corcho, *Ontology Engineering Group - Polytechnic University of Madrid, SPAIN*
- Gabriela Csurka, *Xerox Research Centre Europa, FRANCE*
- Mathieu d'Aquin, *Knowledge Media institute (KMi) - Open University, UK*
- Ahmed A. A. Esmin, *Federal University of Lavras, BRAZIL*
- Frédéric Flouvat, *PPME Labs - University of New Caledonia, NEW CALEDONIA*
- Hakim Hacid, *Alcatel-Lucent, Bell Labs, FRANCE*
- Dino Ienco, *Irstea, FRANCE*
- Clement Jonquet, *LIRMM - University of Montpellier 2, FRANCE*
- Eric Kergosien, *Irstea, FRANCE*
- Pierre-Nicolas Mougel, *Hubert Curien Labs - University of Saint-Etienne, FRANCE*
- Phan Nhat Hai, *Oregon State University, USA*
- Thomas Opitz, *University of Montpellier 2 - LIRMM, FRANCE*
- Jordi Nin, *Barcelona Supercomputing Center (BSC) - Technical University of Catalonia (BarcelonaTECH), SPAIN*
- Gabriela Pasi, *Information Retrieval Lab - University of Milan Bicocca, ITALY*
- Miguel Nuñez del Prado Cortez, *Intersec Labs - Paris, FRANCE*
- José Manuel Perea Ortega, *European Commission - Joint Research Centre (JRC)- Ispra, ITALY*
- Yoann Pitarch, *IRIT - Toulouse, FRANCE*
- Pascal Poncelet, *LIRMM - University of Montpellier 2, FRANCE*
- Julien Rabatel, *LIRMM, FRANCE*
- Mathieu Roche, *Cirad - TETIS and LIRMM, FRANCE*
- Nancy Rodriguez, *LIRMM - University of Montpellier 2, FRANCE*
- Hassan Saneifar, *Raja University, IRAN*
- Nazha Selmaoui-Folcher, *PPME Labs - University of New Caledonia, NEW CALEDONIA*
- Maguelonne Teisseire, *Irstea - LIRMM, FRANCE*
- Boris Villazon-Terrazas, *iLAB Research Center - iSOCO, SPAIN*
- Pattaraporn Warintarawej, *Prince of Songkla University, THAILAND*
- Osmar Zaïane, *Department of Computing Science, University of Alberta, CANADA*

SIMBig 2014 Paper Contents

SIMBig 2014 Keynote Speakers

Mining Social Networks: Challenges and Directions

Pascal Poncelet, Lirmm, Montpellier, France

Professor Pascal Poncelet talk was focused on analysis of data associate to social network. In his presentation, Professor Poncelet summarized several techniques through knowledge discovery on databases process.

NLP approaches: How to Identify Relevant Information in Social Networks?

Mathieu Roche, UMR TETIS, Irstea, France

In this talk, doctor Roche outlined last tendencies in text mining task. Several techniques (sentiment analysis, opinion mining, entity recognition, ...) on different corpus (tweets, blogs, sms,...) were detailed in this presentation.

Social Network Analysis: Overview and Applications

Hakim Hacid, Zayed University, Dubai, UAE

Doctor Hacid presented a complete overview concerning techniques associated to explote social web data. Techniques as social network analysis, social information retrieval and mashups full completion were presented.

Putting intelligence in Web Data With Examples Education

Mathieu d’Aquin, Knowledge Media Institute, The Open University, UK

Doctor d'Aquin presented several web mining approaches addressed to education issues. In this talk doctor d'Aquin detailed themes like intelligent web information and knowledge processing, the semantic web, among others.

Identification of Opinion Leaders Using Text Mining Technique in Virtual Community

Chihli Hung
Department of Information Management
Chung Yuan Christian University
Taiwan 32023, R.O.C.
chihli@cycu.edu.tw

Pei-Wen Yeh
Department of Information Management
Chung Yuan Christian University
Taiwan 32023, R.O.C.
moguflyegmail.com

Abstract

Word of mouth (WOM) affects the buying behavior of information receivers stronger than advertisements. Opinion leaders further affect others in a specific domain through their new information, ideas and opinions. Identification of opinion leaders has become one of the most important tasks in the field of WOM mining. Existing work to find opinion leaders is based mainly on quantitative approaches, such as social network analysis and involvement. Opinion leaders often post knowledgeable and useful documents. Thus, the contents of WOM are useful to mine opinion leaders as well. This research proposes a text mining-based approach to evaluate features of expertise, novelty and richness of information from contents of posts for identification of opinion leaders. According to experiments in a real-world bulletin board data set, this proposed approach demonstrates high potential in identifying opinion leaders.

1 Introduction

This research identifies opinion leaders using the technique of text mining, since the opinion leaders affect other members via word of mouth (WOM) on social networks. WOM defined by Arndt (1967) is an oral person-to-person communication means between an information receiver and a sender, who exchange the experiences of a brand, a product or a service based on a non-commercial purpose. Internet provides human beings with a new way of communication. Thus, WOM influences the consumers more quickly, broadly, widely, significantly and consumers are further influenced by other consumers without any geographic limitation (Flynn et al., 1996).

Nowadays, making buying decisions based on WOM becomes one of collective decision-making strategies. It is nature that all kinds of human groups have opinion leaders, explicitly or implicitly (Zhou et al., 2009). Opinion leaders usually have a stronger influence on other members through their new information, ideas and representative opinions (Song et al., 2007). Thus, how to identify opinion leaders has increasingly attracted the attention of both practitioners and researchers.

As opinion leadership is relationships between members in a society, many existing opinion leader identification tasks define opinion leaders by analyzing the entire opinion network in a specific domain, based on the technique of social network analysis (SNA) (Kim, 2007; Kim and Han, 2009). This technique depends on relationship between initial publishers and followers. A member with the greatest value of network centrality is considered as an opinion leader in this network (Kim, 2007).

However, a junk post does not present useful information. A WOM with new ideas is more interesting. A spam link usually wastes readers' time. A long post is generally more useful than a short one (Agarwal et al., 2008). A focused document is more significant than a vague one. That is, different documents may contain different influences on readers due to their quality of WOM. WOM documents per se can also be a major indicator for recognizing opinion leaders. However, such quantitative approaches, i.e. number-based or

SNA-based methods, ignore quality of WOM and only include quantitative contributions of WOM.

Expertise, novelty, and richness of information are three important features of opinion leaders, which are obtained from WOM documents (Kim and Han, 2009). Thus, this research proposes a text mining-based approach in order to identify opinion leaders in a real-world bulletin board system.

Besides this section, this paper is organized as follows. Section 2 gives an overview of features of opinion leaders. Section 3 describes the proposed text mining approach to identify opinion leaders. Section 4 describes the data set, experiment design and results. Finally, a conclusion and further research work are given in Section 5.

2 Features of Opinion Leaders

The term "opinion leader", proposed by Katz and Lazarsfeld (1957), comes from the concept of communication. Based on their research, the influence of an advertising campaign for political election is lesser than that of opinion leaders. This is similar to findings in product and service markets. Although advertising may increase recognition of products or services, word of mouth disseminated via personal relations in social networks has a greater influence on consumer decisions (Arndt, 1967; Khammash and Griffiths, 2011). Thus, it is important to identify the characteristics of opinion leaders.

According to the work of Myers and Robertson (1972), opinion leaders may have the following seven characteristics. Firstly, opinion leadership in a specific topic is positively related to the quantity of output of the leader who talks, knows and is interested in the same topic. Secondly, people who influence others are themselves influenced by others in the same topic. Thirdly, opinion leaders usually have more innovative ideas in the topic. Fourthly and fifthly, opinion leadership is positively related to overall leadership and an individual's social leadership. Sixthly, opinion leaders usually know more about demographic variables in the topic. Finally, opinion leaders are domain dependent. Thus, an opinion leader influences others in a specific topic in a social network. He or she knows more about this topic and publishes more new information.

Opinion leaders usually play a central role in a social network. The characteristics of typical network hubs usually contain six aspects, which are ahead in adoption, connected, travelers, information-hungry, vocal, and exposed to media more than others (Rosen, 2002). Ahead in adoption means that network hubs may not be the first to adopt new products but they are usually ahead of the rest in the network. Connected means that network hubs play an influential role in a network, such as an information broker among various different groups. Traveler means that network hubs usually love to travel in order to obtain new ideas from other groups. Information-hungry means that network hubs are expected to provide answers to others in their group, so they pursue lots of facts. Vocal means that network hubs love to share their opinions with others and get responses from their audience. Exposed to media means that network hubs open themselves to more communication from mass media, and especially to print media. Thus, a network hub or an opinion leader is not only an influential node but also a novelty early adopter, generator or spreader. An opinion leader has rich expertise in a specific topic and loves to be involved in group activities.

As members in a social network influence each other, degree centrality of members and involvement in activities are useful to identify opinion leaders (Kim and Han, 2009). Inspired by the PageRank technique, which is based on the link structure (Page et al., 1998), OpinionRank is proposed by Zhou et al. (2009) to rank members in a network. Jiang et al. (2013) proposed an extended version of PageRank based on the sentiment analysis and MapReduce. Agarwal et al. (2008) identified influential bloggers through four aspects, which are recognition, activity generation, novelty and eloquence. An influential blog is recognized by others when this blog has a lot of in-links. The feature of activity generation is measured by how many comments a post receives and the number of posts it initiates. Novelty means novel ideas, which may attract many in-links from the blogs of others. Finally, the feature of eloquence is evaluated by the length of post. A lengthy post is treated as an influential post.

Li and Du (2011) determined the expertise of authors and readers according to the similarity between their posts and the pre-built term ontology. However both features of information novelty and influential position are dependent on linkage relationships between blogs. We propose a novel text mining-based approach and compare it with several quantitative approaches.

3 Quality Approach-Text Mining

Contents of word of mouth contain lots of useful information, which has high relationships with important features of opinion leaders. Opinion leaders usually provide knowledgeable and novel information in their posts (Rosen, 2002; Song et al., 2007). An influential post is often eloquent (Keller and Berry, 2003). Thus, expertise, novelty, and richness of information are important characteristics of opinion leaders.

3.1 Preprocessing

This research uses a traditional Chinese text mining process, including Chinese word segmenting, part-of-speech filtering and removal of stop words for the data set of documents. As a single Chinese character is very ambiguous, segmenting Chinese documents into proper Chinese words is necessary (He and Chen, 2008). This research uses the CKIP service (http://ckipsvr.iis.sinica.edu.tw/) to segment Chinese documents into proper Chinese words and their suitable part-of-speech tags. Based on these processes, 85 words are organized into controlled vocabularies as this approach is efficient to capture the main concepts of document (Gray et al., 2009).

3.2 Expertise

This can be evaluated by comparing their posts with the controlled vocabulary base (Li and Du, 2011). For member *i*, words are collected from his or her posted documents and member vector *i* is represented as $f_i=(w_1, w_2, \ldots w_j, \ldots, w_N)$, where w_j denotes the frequency of word *j* used in the posted documents of user *i*. *N* denotes the number of words in the controlled vocabulary. We then normalize the member vector by his or her maximum frequency of any significant word. The degree of expertise can be calculated by the Euclidean norm as show in (1).

$$\exp_i = \left\| \frac{f_i}{m_i} \right\|, \tag{1}$$

where $\|\bullet\|$ is Euclidean norm.

3.3 Novelty

We utilize Google trends service (http://www.google.com/trends) to obtain the first-search time tag for significant words in documents. Thus, each significant word has its specific time tag taken from the Google search repository. For example, the first-search time tag for the search term, Nokia N81, is 2007 and for Nokia Windows Phone 8 is 2011. We define three degrees of novelty evaluated by the interval between the first-search year of significant words and the collected year of our targeted document set, i.e. 2010. This significant word belongs to normal novelty if the interval is equal to two years. A significant word with an interval of less than two years belongs to high novelty and one with an interval greater than two years belongs to low novelty. We then summarize all novelty values based on significant words used by a member in a social network. The equation of novelty for a member is shown in (2).

$$nov_i = \frac{e_h + 0.66 \times e_m + 0.33 \times e_l}{e_h + e_m + e_l}, \tag{2}$$

where e_h , e_m and e_l is the number of words that belong to the groups of high, normal and low novelty, respectively.

3.4 Richness of Information

In general, a long document suggests some useful information to the users (Agarwal et al., 2008). Thus, richness of information of posts can be used for the identification of opinion leaders. We use both textual information and multimedia information to represent the richness of information as (3).

$$ric=d + g, \tag{3}$$

where *d* is the total number of significant words that the user uses in his or her posts and *g* is the total number of multimedia objects that the user posts.

3.5 Integrated Text Mining Model

Finally, we integrate expertise, novelty and richness of information from the content of posted documents. As each feature has its own distribution and range, we normalize each feature to a value between 0 and 1. Thus, the weights of opinion leaders based on the quality of posts become the average of these three features as (4).

$$ITM = \frac{Norm(nov) + Norm(exp) + Norm(ric)}{3}. \quad (4)$$

4 Experiments

4.1 Data Set

Due to lack of available benchmark data set, we crawl WOM documents from the Mobile01 bulletin board system (http://www.mobile01.com/), which is one of the most popular online discussion forums in Taiwan. This bulletin board system allows its members to contribute their opinions free of charge and its contents are available to the public. A bulletin board system generally has an organized structure of topics. This organized structure provides people who are interested in the same or similar topics with an online discussion forum that forms a social network. Finding opinion leaders on bulletin boards is important since they contain a lot of availably focused WOM. In our initial experiments, we collected 1537 documents, which were initiated by 1064 members and attracted 9192 followers, who posted 19611 opinions on those initial posts. In this data set, the total number of participants is 9460.

4.2 Comparison

As we use real-world data, which has no ground truth about opinion leaders, a user centered evaluation approach should be used to compare the difference between models (Kritikopoulos et al., 2006). In our research, there are 9460 members in this virtual community. We suppose that ten of them have a high possibility of being opinion leaders.

As identification of opinion leaders is treated to be one of important tasks of social network analysis (SNA), we compare the proposed model (i.e. ITM) with three famous SNA approaches, which are degree centrality (DEG), closeness centrality (CLO), betweenness centrality (BET). Involvement (INV) is an important characteristic of opinion leaders (Kim and Han, 2009). The number of documents that a member initiates plus the number of derivative documents by other members is treated as involvement.

Thus, we have one qualitative model, i.e. ITM, and four quantitative models, i.e. DEG, CLO, BET and INV. We put top ten rankings from each model in a pool of potential opinion leaders. Duplicate members are removed and 25 members are left. We request 20 human testers, which have used and are familiar with Mobile01.

In our questionnaire, quantitative information is provided such as the number of documents that the potential opinion leaders initiate and the number of derivative documents that are posted by other members. For the qualitative information, a maximum of three documents from each member are provided randomly to the testers. The top 10 rankings are also considered as opinion leaders based on human judgment.

4.3 Results

We suppose that ten of 9460 members are considered as opinion leaders. We collect top 10 ranking members from each models and remove duplicates. We request 20 human testers to identify 10 opinion leaders from 25 potential opinion leaders obtained from five models. According to experiment results in Table 1, the proposed model outperforms others. This presents the significance of documents per se. Even INV is a very simple approach but it performs much better than social network analysis models, i.e. DEG, CLO and BET. One possible reason is the sparse network structure. Many sub topics are in the bulletin board system so these topics form several isolated sub networks.

	Recall	Precision	*F*-measure	Accuracy
DEG	0.45	0.50	0.48	0.56
CLO	0.36	0.40	0.38	0.48
BET	0.64	0.70	0.67	0.72
INV	0.73	0.80	0.76	0.80
ITM	0.82	0.90	0.86	0.88

Table 1: Results of models evaluated by recall, precision, F-measure and accuracy

5 Conclusions and Further Work

Word of mouth (WOM) has a powerful effect on consumer behavior. Opinion leaders have stronger influence on other members in an opinion society. How to find opinion leaders has been of interest to both practitioners and researchers. Existing models mainly focus on quantitative features of opinion leaders, such as the number of posts and the central position in the social network. This research considers this issue from the viewpoints of text mining. We propose an integrated text mining model by extracting three important features of opinion leaders regarding novelty, expertise and richness of information, from documents. Finally, we compare this proposed text mining model with four quantitative approaches, i.e., involvement, degree centrality, closeness centrality and betweenness centrality, evaluated by human judgment. In our experiments, we found that the involvement approach is the best one among the quantitative approaches. The text mining approach outperforms its quantitative counterparts as the richness of document information provides a similar function to the qualitative features of opinion leaders. The proposed text mining approach further measures opinion leaders based on features of novelty and expertise.

In terms of possible future work, some integrated strategies of both qualitative and quantitative approaches should take advantages of both approaches. For example, the 2-step integrated strategy, which uses the text mining-based approach in the first step, and uses the quantitative approach based on involvement in the second step, may achieve the better performance. Larger scale experiments including topics, the number of documents and testing, should be done further in order to produce more general results.

References

Agarwal, N., Liu, H., Tang, L. and Yu, P. S. 2008. Identifying the Influential Bloggers in a Community. Proceedings of WSDM, 207-217.

Arndt, J. 1967. Role of Product-Related Conversations in the Diffusion of a New Product. Journal of Marketing Research, 4(3):291-295.

Flynn, L. R., Goldsmith, R. E. and Eastman, J. K. 1996. Opinion Leaders and Opinion Seekers: Two New Measurement Scales. Academy of Marketing

He, J. and Chen, L. 2008. Chinese Word Segmentation Based on the Improved Particle Swarm Optimization Neural Networks. Proceedings of IEEE Cybernetics and Intelligent Systems, 695-699.

Jiang, L., Ge, B., Xiao, W. and Gao, M. 2013. BBS Opinion Leader Mining Based on an Improved PageRank Algorithm Using MapReduce. Proceedings of Chinese Automation Congress, 392-396.

Katz, E. and Lazarsfeld, P. F. 1957. Personal Influence, New York: The Free Press.

Keller, E. and Berry, J. 2003. One American in Ten Tells the Other Nine How to Vote, Where to Eat and, What to Buy. They Are The Influentials. The Free Press.

Khammash, M. and Griffiths, G. H. 2011. Arrivederci CIAO.com Buongiorno Bing.com- Electronic Word-of-Mouth (eWOM), Antecedences and Consequences. International Journal of Information Management, 31:82-87.

Kim, D. K. 2007. Identifying Opinion Leaders by Using Social Network Analysis: A Synthesis of Opinion Leadership Data Collection Methods and Instruments. PhD Thesis, the Scripps College of Communication, Ohio University.

Kim, S. and Han, S. 2009. An Analytical Way to Find Influencers on Social Networks and Validate their Effects in Disseminating Social Games. Proceedings of Advances in Social Network Analysis and Mining, 41-46.

Kritikopoulos, A., Sideri, M. and Varlamis, I. 2006. BlogRank: Ranking Weblogs Based on Connectivity and Similarity Features. Proceedings of the 2nd International Workshop on Advanced Architectures and Algorithms for Internet Delivery and Applications, Article 8.

Li, F. and Du, T. C. 2011. Who Is Talking? An Ontology-Based Opinion Leader Identification Framework for Word-of-Mouth Marketing in Online Social Blogs. Decision Support Systems, 51, 2011:190-197.

Myers, J. H. and Robertson, T. S. 1972. Dimensions of Opinion Leadership. Journal of Marketing Research, 4:41-46.

Page, L., Brin, S., Motwani, R. and Winograd, T. 1998. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report, Stanford University.

Rosen, E. 2002. The Anatomy of Buzz: How to Create Word of Mouth Marketing, 1st ed., Doubleday.

Song, X., Chi, Y., Hino, K. and Tseng, B. L. 2007. Identifying Opinion Leaders in the Blogosphere. Proceedings of CIKM'07, 971-974.

Zhou, H., Zeng, D. and Zhang, C. 2009. Finding Leaders from Opinion Networks. Proceedings of the 2009 IEEE International Conference on Intelligence and Security Informatics, 266-268.

Quality Metrics for Optimizing Parameters Tuning in Clustering Algorithms for Extraction of Points of Interest in Human Mobility

Miguel Nuẽz del Prado Cortez
Peru I+D+I
Technopark IDI
miguel.nunez@peruidi.com

Hugo Alatrista-Salas
GRPIAA Labs., PUCP
Peru I+D+I
halatrista@pucp.pe

Abstract

Clustering is an unsupervised learning technique used to group a set of elements into non-overlapping clusters based on some predefined dissimilarity function. In our context, we rely on clustering algorithms to extract points of interest in human mobility as an inference attack for quantifying the impact of the privacy breach. Thus, we focus on the input parameters selection for the clustering algorithm, which is not a trivial task due to the direct impact of these parameters in the result of the attack. Namely, if we use too relax parameters we will have too many point of interest but if we use a too restrictive set of parameters, we will find too few groups. Accordingly, to solve this problem, we propose a method to select the best parameters to extract the optimal number of POIs based on quality metrics.

1 Introduction

The first step in inference attacks over mobility traces is the extraction of the point of interest (POI) from a trail of mobility traces. Indeed, this phase impacts directly the global accuracy of an inference attack that relies on POI extraction. For instance, if an adversary wants to discover Alice's home and place of work the result of the extraction must be as accurate as possible, otherwise they can confuse or just not find important places. In addition, for a more sophisticated attack such as next place prediction, a mistake when extracting POIs can decrease significantly the global precision of the inference. Most of the extraction techniques use heuristics and clustering algorithms to extract POIs from location data. On one hand, heuristics rely on the *dwell time*, which is the lost of signal when user gets into a building. Another used heuristic is the *residence time*, which represents the time that a user spends at a particular place. On the other hand, clustering algorithms group nearby mobility traces into clusters.

In particular, in the context of POI extraction, it is important to find a suitable set of parameters~~,~~ for a specific cluster algorithm~~,~~ in order to obtain a good accuracy as result of the clustering. The main contribution of this paper is a methodology to find a "optimal" configuration of input parameters for a clustering algorithm based on quality indices. This optimal set of parameters allows us to have the appropriate number of POIs in order to perform another inference attack. This paper is organized as follows. First, we present some related works on parameters estimation techniques in Section 2. Afterwards, we describe the clustering algorithms used to perform the extraction of points of interests (POIs) as well as the metrics to measure the quality of formed clusters in sections 3 and 4, respectively. Then, we introduce the method to optimize the choice of the parameters in Section 5. Finally, Section 6 summarizes the results and presents the future directions of this paper.

2 Related works

Most of the previous works estimate the parameters of the clustering algorithms for the point of interest extraction by using empirical approaches or highly computationally expensive methods. For instance, we use for illustration purpose two classical clustering approaches, K-means (MacQueen et al., 1967) and DBSCAN (Ester et al., 1996). In the former clustering algorithm, the main issue is how to determine k, the number of clusters. Therefore, several approaches have been proposed to address this issue (Hamerly and Elkan, 2003; Pham et al., 2005). The latter algorithm relies on OPTICS (Ankerst et al., 1999) algorithm, which searches the space of parameters of DBSCAN in order to find the optimal number of clusters. The more parameters the clustering algorithm has, the bigger the combinatorial space of parameters is. Nevertheless, the methods to calibrate cluster algorithm inputs do not guarantee a good accuracy for extracting meaningful POIs. In the next section, we described the cluster algorithms used in our study.

3 Clustering algorithms for extraction of points of interest

To perform the POI extraction, we rely on the following clustering algorithms:

3.1 Density Joinable Cluster (DJ-Cluster)

DJ-Cluster (Zhou et al., 2004) is a clustering algorithm taking as input a minimal number of points $minpts$, a radius r and a trail of mobility traces M. This algorithm works in two phases. First, the pre-processing phase discards all the moving points (*i.e.* whose speed is above ϵ, for ϵ a small value) and then, squashes series of repeated static points into a single occurrence for each series. Next, the second phase clusters the remaining points based on neighborhood density. More precisely, the number of points in the neighborhood must be equal or greater than $minpts$ and these points must be within radius r from the medoid of a set of points. Where medioid is the real point *m* that minimizes the sum of distances from the point *m* to the other points in the cluster. Then, the algorithm merges the new cluster with the clusters already computed, which share at least one common point. Finally, during the merging, the algorithm erases old computed clusters and only keeps the new cluster, which contains all the other merged clusters.

3.2 Density Time Cluster (DT-Cluster)

DT-Cluster (Hariharan and Toyama, 2004) is an iterative clustering algorithm taking as input a distance threshold d, a time threshold t and a trail of mobility traces M. First, the algorithm starts by building a cluster C composed of all the consecutive points within distance d from each other. Afterwards, the algorithm checks if the accumulated time of mobility traces between the youngest and the oldest ones is greater than the threshold t. If it is the case, the cluster is created and added to the list of POIs. Finally as a post-processing step, DT-Cluster merges the clusters whose medioids are less than $d/3$ far from each other.

3.3 Time Density (TD-Cluster)

Introduced in (Gambs et al., 2011), TD-Cluster is a clustering algorithm inspired from DT Cluster, which takes as input parameters a radius r, a time window t, a tolerance rate τ, a distance threshold d and a trail of mobility traces M. The algorithm starts by building iteratively clusters from a trail M of mobility traces that are located within the time window t. Afterwards, for each cluster, if a fraction of the points (above the tolerance rate τ) are within radius r from the medoid, the cluster is integrated to the list of clusters outputted, whereas otherwise it is simply discarded. Finally, as for DT Cluster, the algorithm merges the clusters whose medoids are less than d far from each other.

3.4 Begin-end heuristic

The objective of the *begin and end location finder* inference attack (Gambs et al., 2010) is to take as meaningful points the first and last of a journey. More precisely, this heuristic considers that the beginning and ending locations of a user, for each working day, might convey some meaningful information.

Since we have introduced the different clustering algorithms to extract points of interest, we present in the next section the indices to measure the quality of the clusters.

4 Cluster quality indices

One aspect of the extraction of POIs inference attacks is the quality of the obtained clusters, which impacts on the precision and recall of the attack. In the following subsection we describe some metrics to quantify how accurate or “how good“ is the outcome of the clustering task. Intuitively, a good clustering is one that identifies a group of clusters that are well separated one from each other, compact and representative. Table 1 summarizes the notation used in this section.

Symbol	Definition
C	An ensemble of clusters.
c_i	The i^{th} cluster of C.
n_c	The number of clusters in C.
m_i	The medoid point of the i^{th} cluster.
$\mathsf{d}(x, y)$	The Euclidean distance between x and y.
$\vert c_i \vert$	The number of points in a cluster c_i.
m'	The closest point to the medoid m_i.
m''	The second closest point to the medoid m_i.
$\vert C \vert$	The total number of points in a set of C.

Table 1: Summary of notations

4.1 Intra-inter cluster ratio

The *intra-inter* cluster ratio (Hillenmeyer, 2012) measures the relation between compact (Equation 1) and well separated groups (Equation 3). More precisely, we first take the inter-cluster distance, which is the average distance from each point in a cluster c_i to its medoid m_i.

$$DIC(c_i) = \frac{1}{|c_i| - 1} \sum_{x_j \in c_i, x_j \neq m_i}^{|c_i|} d(x_j, m_i) \qquad (1)$$

Then, the average intra-cluster distance (***DIC***) is computed using Equation 2.

$$AVG\_DIC(C) = \frac{1}{n_c} \sum_{c_i \in C}^{|C|} DIC(c_i) \qquad (2)$$

Afterwards, the mean distance among all medoids (***DOC***) in the cluster C is computed, using Equation 3.

$$DOC(C) = \frac{1}{|n_C| \times (|n_C| - 1)} \sum_{c_i \in C}^{|C|} \sum_{c_j \in C, i \neq j}^{|C|} d(m_i, m_j) \qquad (3)$$

Finally, the ratio intra-inter cluster *rii* is given by the Equation 4 as the relationship between the average intra cluster distance divided by the inter-cluster distance.

$$rii(C) = \frac{AVG\_DIC(C)}{DOC(C)} \qquad (4)$$

The intra-inter ratio has an approximate linear complexity in the number of points to be computed and gives low values to well separated and compact cluster.

4.2 Additive margin

Inspired by the Ben-David and Ackerman (Ben-David and Ackerman, 2008) *k-additive Point Margin* (*K-AM*) metric , which evaluates how well centered clusters are. We measure the difference between the medoid m_i and its two closest points m' and m'' of a given group c_i belonging to a cluster C (Equation 5).

$$K - AM(c_i) = d(m_i, m_i'') - d(m_i, m_i') \qquad (5)$$

Since the average of the k-additive point margins for all groups c_i in a cluster C is computed, we take the ratio between the average k-additive Point Margin and the minimal inter-cluster distance (Equation 1) as shown in Equation 6.

$$AM(C) = min_{c_i \in C} \frac{\frac{1}{n_c} \sum_{c_i \in C}^{n_c} K - AM(c_i)}{DIC(c_i)} \qquad (6)$$

The additive margin method has a linear complexity in the number of clusters. This metric gives a high value for a well centered clusters.

4.3 Information loss

The information loss ratio is a metric inspired by the work of Sole and coauthors (Solé et al., 2012). The basic idea is to measure the percent of information that is lost while representing original data only by a certain number of groups (*e.g.*, when we represent the POIs by the cluster medoids instead of the whole set of points). To evaluate the percent of information loss, we compute the sum of distance of each point represented by x_i to its medoid m_i for all clusters $c_i \in C$ as we shown in Equation 7.

$$SSE(C) = \sum_{c_i \in C}^{n_c} \sum_{x_j \in c_i}^{|c|} d(x_j, m_i) \qquad (7)$$

Then, we estimate the accumulated distance of all points of a trail of mobility traces in the cluster C to a global centroid (GC) using the following equation Equation 8.

$$SST(C) = \sum_{x_i \in C}^{|C|} d(x_i, GC) \qquad (8)$$

Finally, the ratio between aforementioned distances is computed using Equation 9, which results in the information loss ratio.

$$IL(C) = \frac{SSE(C)}{SST(C)} \tag{9}$$

The computation of this ratio has a linear complexity. The lowest is the value of this ratio, the more representative the clusters are.

4.4 Dunn index

This quality index (Dunn, 1973; Halkidi et al., 2001) attempts to recognize compact and well-separated clusters. The computation of this index relies on a dissimilarity function (*e.g.* Euclidean distance d) between medoids and the diameter of a cluster (*c.f*, Equation 10) as a measure of dispersion.

$$diam(c_i) = max_{x,y \in c_i, x \neq y} d(x, y) \tag{10}$$

Then, if the clustering C is compact (*i.e*, the diameters tend to be small) and well separated (distance between cluster medoids are large), the result of the index, given by the Equation 11, is expected to be large.

$$\begin{aligned} DIL(C) = min_{c_i \in C}[min_{c_j \in C, j=i+1} \\ [\frac{d(m_i, m_j)}{max_{c_k \in C} diam(c_k)}]] \end{aligned} \tag{11}$$

The greater is this index, the better the performance of the clustering algorithm is assumed to be. The main drawbacks of this index is the computational complexity and the sensitivity to noise in data.

4.5 Davis-Bouldin index

The objective of the Davis-Bouldin index (*DBI*) (Davies and Bouldin, 1979; Halkidi et al., 2001) is to evaluate how well the clustering was performed by using properties inherent to the dataset considered. First, we use a scatter function within the cluster c_i of the clustering C (Equation 12).

$$S(c_i) = \sqrt{\frac{1}{n_c} \sum_{x_j \in c_i}^{n} d(x_j, m_i)^2} \tag{12}$$

Then, we compute the distance between two different clusters c_i and c_j, given by Equation 13.

$$M(c_i, c_j) = \sqrt{d(m_i, m_j)} \tag{13}$$

Afterwards, a similarity measure between two clusters c_i and c_j, called *R*-similarity, is estimated, based on Equation 14.

$$R(c_i, c_j) = \frac{S(c_i) + S(c_j)}{M(c_i, c_j)} \tag{14}$$

After that, the most similar cluster c_j to c_i is the one maximizing the result of the function $R_{all}(c_i)$, which is given by Equation 15 for $i \neq j$.

$$R_{all}(c_i) = max_{c_j \in C, i \neq j} R(c_i, c_j) \tag{15}$$

Finally, the DBI is equal to the average of the similarity between clusters in the clustering set C (Equation 16).

$$DBI(C) = \frac{1}{n_c} \sum_{c_i \in C}^{n_c} R_{all}(c_i) \tag{16}$$

Ideally, the clusters $c_i \in C$ should have the minimum possible similarity to each other. Accordingly, the lower is the DB index, the better is the clustering formed. These indices would be used to maximize the number of significant places a cluster algorithm could find. More precisely, in the next section we evaluate the cluster algorithm aforementioned as well as the method to extract the meaningful places using the quality indices.

5 Selecting the optimal parameters for clustering

In order to establish how to select the best set of parameters for a given clustering algorithm, we have computed the precision, recall and F-measure of all users of LifeMap dataset (Chon and Cha, 2011). One of the unique characteristic of this dataset is that the POIs have been annotated by the users. Consequently, given a set of clusters $c_i \in C$ such that $C = \{c_1, c_2, c_3, \ldots, c_n\}$ and a set of points of interest (POIs) defined by the users $P_{poi} = \{p_{poi\ 1}, p_{poi\ 2}, p_{poi\ 3}, \ldots, p_{poi\ n}\}$ we were able to compute the precision, recall and f-measure as we detail in the next subsection.

5.1 Precision, recall and F-measure

To compute the recall (*c.f.* Equation 17), we take as input a clustering set C, the *ground truth* represented by the vector P_{poi} (which was defined manually by each user) as well as a $radius$ to count all the clusters $c \in C$ that are within the $radius$ of $p_{poi} \in P_{poi}$, which represents the "good clusters". Then, the ratio of the number of *good clusters* compared to the *total number of found clusters* is computed. This measure illustrates the ratio of extracted cluster that are POIs divided by the total number of extracted clusters.

$$Precision = \frac{good\ clusters}{total\ number\ extracted\ clusters} \quad (17)$$

To compute the recall (*c.f.* Equation 18), we take as input a clustering set C, a vector of POIs P_{poi} as well as a $radius$ to count the discovered POIs $p_{poi} \in P_{poi}$ within a $radius$ of the clusters $c \in C$, which represents the "good POIs". Then, the ratio between the number of *good POIs* and the *total number of POIs* is evaluated. This metric represents the percent of the extracted unique POIs.

$$Recall = \frac{good\ POIs}{total\ number\ of\ POIs} \quad (18)$$

Finally, the F-measure is defined as the weighted average of the precision and recall as we can see in Equation 19.

$$F - measure = 2 \times \frac{precision \times recall}{precision + recall} \quad (19)$$

We present the dataset used for our experiments in the next subsection.

5.2 Dataset description

In order to evaluate our approach, we use the *LifeMap dataset* (Chon et al., 2012), which is composed of mobility traces of 12 user collected for a year in Seoul, Korea. This dataset comprises location (latitude and longitude) collected with a frequency between 2 to 5 minutes with the user defined point of interest as *true* if the mobility trace is considered as important or meaningfull for each user. Table 2 summarizes the main characteristics of this dataset, such as the collect period, the average number of traces per user, the total number of mobility traces in the dataset, the minimal and maximal number of users' mobility traces.

Since we have described our dataset, we present the results of our experiments in the next subsection.

Characteristics	LifeMap
Total nb of users	12
Collection period (nb of days)	366
Average nb of traces/user	4 224
Total nb of traces	50 685
Min #Traces for a user	307
Max #Traces for a user	9 473

Table 2: Main characteristics of the LifeMap dataset.

5.3 Experimental results

This section is composed of two parts, in the first part we compare the performance of the previously described clustering algorithms, with two baseline clustering algorithms namely *k*-means and DBSCAN. In the second part, a method to select the most suitable parameters for a clustering algorithm is presented.

Input parameters	Possible values	DBSCAN	DJ cluster	DT cluster	K-means	TD cluster
Tolerance rate (%)	{0.75, 0.8, 0.85, 0.9}	y	Y	y	y	y
Tolerance rate (%)	{0.75, 0.8, 0.85, 0.9}	✗	✗	✗	✗	✓
Minpts (points)	{3, 4, 5, 6, 7, 8, 9, 10, 20, 50}	✓	✓	✗	✗	✗
Eps (Km.)	{0.01, 0.02, 0.05, 0.1, 0.2}	✓	✓	✓	✗	✓
Merge distance (Km.)	{0.02, 0.04, 0.1, 0.2, 0.4}	✗	✗	✓	✗	✓
Time shift (hour)	{1, 2, 3, 4, 5, 6}	✗	✗	✓	✗	✓
K (num. clusters)	{5, 6, 7,8, 9}	✗	✗	✗	✓	✗

Table 3: Summary of input parameters for clustering algorithms.

	Precision	Recall	F-measure	Time(s)	Number of parameters	Complexity
DBSCAN	0,58	0,54	0,48	316	3	$O(n^2)$
DJ-Cluster	0,74	0,52	0,52	429	3	$O(n^2)$
DT-Cluster	0,38	0,47	0,39	279	3	$O(n^2)$
k-means	0,58	0,51	0,49	299	2	$O(n)$
TD-Cluster	0,43	0,54	0,44	362	4	$O(n^2)$

Table 4: The characteristics of the clustering algorithms.

In order to compare the aforementioned clustering algorithms, we have take into account the precision, recall, F-measure obtained, average execution time, number of input parameters and time complexity. To evaluate these algorithms, we used the LifeMap dataset with POIs annotation and a set of different parameters configurations for each algorithm, which are summarized in Table 3. After running these configurations, we obtained the results shown in Table 4 for the different input values.

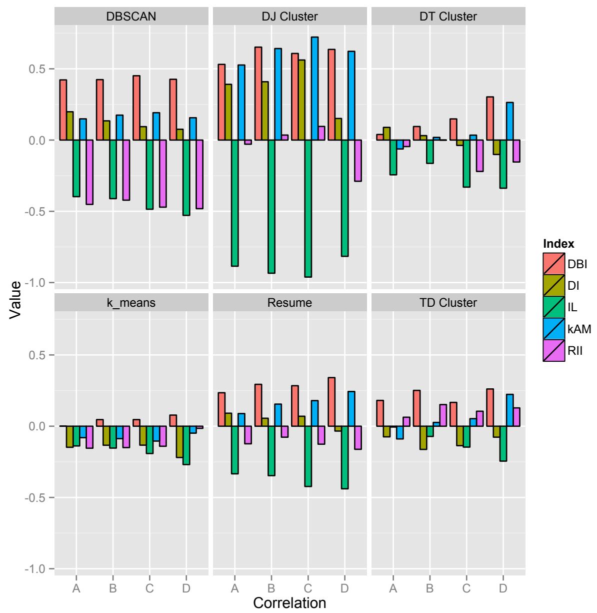


Figure 1: Correlation of quality indices with the computed F-measure. Where A) is the correlation measured between the user annotation and the centroid at 20 m of radius B) at 35 m radius C) at 50 m radius, D) at 100 m radius and DBI=Davis-Bouldin index, DI=Dunn index, IL=Information loss, kAM=Additive margin and RII= Ratio intra-inter cluster.

It is possible to observe that the precision of DJ-Cluster out performs better than the other clustering algorithms. In terms of recall DBSCAN and TD-Cluster perform the best but DJ-Cluster is just behind them. Moreover, DJ-Cluster has the best F-measure. Regarding the execution time, DT-Clustering the fastest one while DJ-Cluster is the slowest algorithm due to the preprocessing phase. Despite the high computational time of DJ-Cluster, this algorithm performs well in terms of F-measure.

In the following, we describe our method to choose "optimal" parameters for obtaining a good F-measure. We have used the aforementioned algorithms with a different set of input parameters configurations for users with POIs annotations in the LifeMap dataset (Chon and Cha, 2011). Once clusters are built, we evaluate the clusters issued from different configurations of distinct algorithms using the previously described quality indices. Afterwards, we were able to estimate the precision, recall and F-measure using the manual annotation of POIs by the users in the LifeMap dataset.

Regarding the relation between the quality indices and the F-measure, we studied the relationship between these factors, in order to identify the indices that are highly correlated with the F-measure, as can be observed in Figure 1. We observe that the two best performing indices, except for k-means, are IL and DBI. The former shows a negative correlation with respect to the F-measure. While the latter, has a positive dependency to the F-measure. Our main objective is to be able to identify the relationship between quality and F-measure among the previous evaluated clustering algorithms. Accordingly, we discard the inter-intra cluster ratio (RII) and the adaptive margin (AM), which only perform well when using k-means and the DJ clustering algorithms. Finally, we observe that the Dunn index has a poor performance. Based on these observations, we were able to propose an algorithm to automatically choose the best configuration of input parameters.

5.4 Parameter selection method

Let us define a vector of parameters $p_i \in P$ and P a set of vectors, such that $P = \{p_1, p_2, p_3, \dots, p_n\}$, a trail of mobility traces M of a user. From previous sections we have the clustering function $C(p_i)$ and the quality metrics Information Loss $IL(C)$ and Davis-Bouldin index $DBI(C)$. Thus, for each vector of parameters we have a tuple composed of the trail of mobility traces, the result of the clustering algorithm and the quality metrics $(p_i, M, C_{p_i}, IL_{C_{p_i}}, DBI_{C_{p_i}})$. When we compute the clustering algorithm and the quality metrics for each vector of parameter for a given user u. We define also a χ'_u matrix, which the matrix χ_u sorted by IL ascending. Finally, the result matrix χ_u is of the form:

$$\chi_u = \begin{vmatrix} p_1 & M & C_{p_1} & IL_{C_{p_1}} & DBI_{C_{p_1}} \\ p_2 & M & C_{p_2} & IL_{C_{p_2}} & DBI_{C_{p_2}} \\ p_3 & M & C_{p_3} & IL_{C_{p_3}} & DBI_{C_{p_3}} \\ \dots & & & & \\ p_n & M & C_{p_n} & IL_{C_{p_n}} & DBI_{C_{p_n}} \end{vmatrix}$$

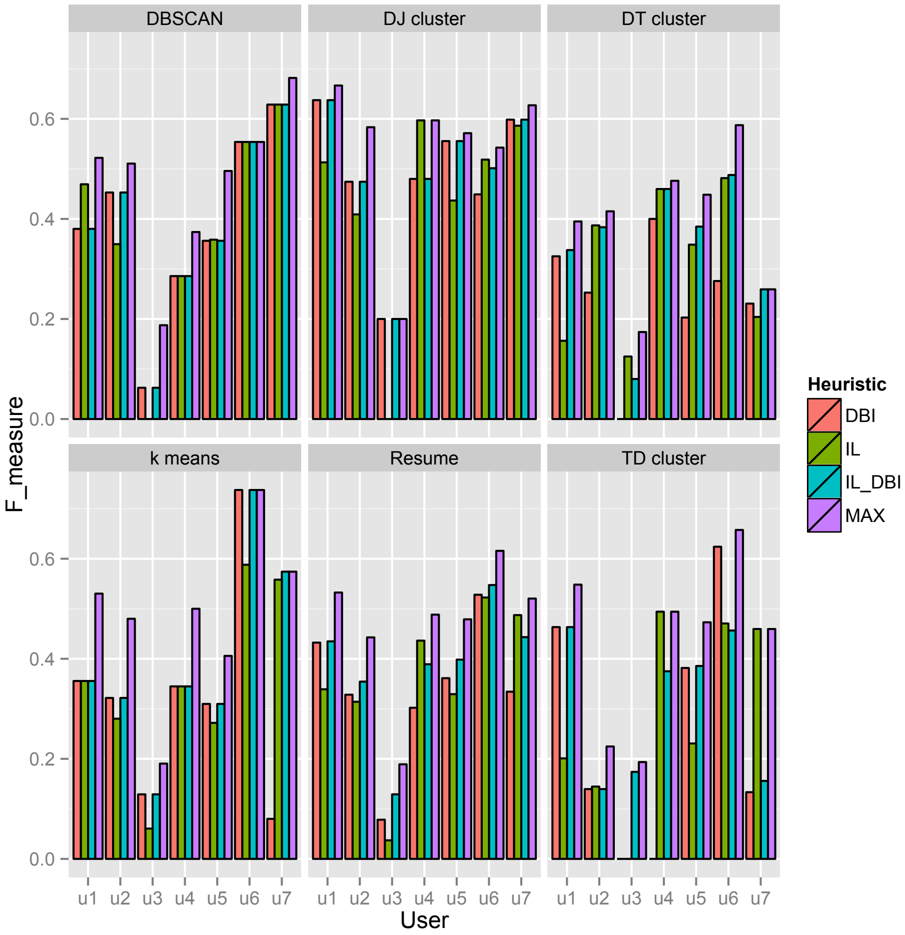


Figure 2: Comparison between F-measure and parameters selection based on schema in Figure **??**. Where DBI=Davis-Bouldin index, IL=Information loss, IL_DBI= combination of IL and DBI and MAX is the maximal computed F-measure (taken as reference to compare with IL_DBI). Resume is the average of all the results using different clustering algorithms.

Therefore, the parameter selection function $S(\chi_u)$ could be defined as:

$$S(\chi_u) = \begin{cases} p_i, & \text{if } max_{p_i}(DBI) \text{ \& } min_{p_i}(IL) \\ p_i', & \text{if } max_{p_i'}(DBI) \text{ in 1st quartile} \end{cases} \tag{20}$$

In detail, the function S takes as input a χ matrix containing the parameters vector p_i, a trail of mobility traces M, the computed clustering $C(p_i, M)$ as well as the quality metrics, such as Information loss $(IL(C))$ and the Davis-Bouldin index $(DBI(C))$. Once all these values have been computed for each evaluated set of parameters, two cases are possible. In the first case, both IL and DBI agree on the same set of input parameters. In the second situation, both IL and DBI refer each one to a different set of parameters. In this case, the algorithm sorts the values by IL in the ascending order (*i.e.*, from the smallest to the largest information loss value). Then, it chooses the set of parameters with the greatest DBI in the first quartile.

For the sake of evaluation, our methodology was tested using the LifeMap dataset to check if the chosen parameters are optimal. We have tested the method with the seven users of LifeMap that have annotated manually their POIs. Consequently, for every set of settings of each clustering algorithm, we have computed the F-measure because we have the ground truth as depicted in Figure 2. The ”MAX” bar represents the best F-measure for the given user and it is compare to the F-measures obtained when using the ”DBI”, ”IL” or ”IL_DBI” as indicators to choose the best input parameters configuration. Finally, this method has a satisfactory performance extracting a good number of POIs for maximizing the F-measure achieving a difference of only 9% with respect to the F-measure computed from the data with the ground truth.

6 Conclusion

In the current paper, we have presented a method to extract the a optimal number of POIs. Consequently, based on the method described in tis paper, we are able to find an appropriate number of POIs relying only on the quality metrics of the extracted clusters and without the knowledge of the ground truth. Nonetheless, we are aware of the small size of dataset but the results encourage us to continue in this direction.

Therefore, in the future we plan to test our method in a larger dataset and in presence of noise like downsamplig or random distortion. Another idea is to evaluate the impact of this method in more complex attacks like prediction of future locations or deanonymization to verify if this step can affect the global result of a chain of inference attacks.

References

Ankerst, M., Breunig, M. M., Kriegel, H.-P., and Sander, J. (1999). Optics: ordering points to identify the clustering structure. *ACM SIGMOD Record*, 28(2):49–60.

Ben-David, S. and Ackerman, M. (2008). Measures of clustering quality: A working set of axioms for clustering. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, pages 121–128, Vancouver, Canada.

Chon, Y. and Cha, H. (2011). LifeMap: A smartphone-based context provider for location-based services. *Pervasive Computing, IEEE*, 10(2):58–67.

Chon, Y., Talipov, E., Shin, H., and Cha, H. (2012). CRAWDAD data set yonsei/lifemap (v. 2012-01-03). Downloaded from http://crawdad.cs.dartmouth.edu/yonsei/lifemap.

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227.

Dunn, J. C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3(3):32–57.

Ester, M., peter Kriegel, H., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Knowledge Discovery and Data Mining*, 2(4):226–231.

Gambs, S., Killijian, M.-O., and Núñez del Prado Cortez, M. (2010). GEPETO: A GEoPrivacy-Enhancing TOolkit. In *Advanced Information Networking and Applications Workshops*, pages 1071–1076, Perth, Australia.

Gambs, S., Killijian, M.-O., and Núñez del Prado Cortez, M. (2011). Show me how you move and I will tell you who you are. *Transactions on Data Privacy*, 2(4):103–126.

Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(3):107–145.

Hamerly, G. and Elkan, C. (2003). Learning the k in K-means. In *In Neural Information Processing Systems*, pages 1–8, Vancouver, Canada.

Hariharan, R. and Toyama, K. (2004). Project lachesis: Parsing and modeling location histories. *Lecture notes in computer science - Geographic information science*, 3(1):106–124.

Hillenmeyer, M. (2012). Intra and inter cluster distance. `http://www.stanford.edu/~maureenh/quals/html/ml/node82.html`.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate mbservations. In *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, Berkeley, CA, USA.

Pham, D. T., Dimov, S. S., and Nguyen, C. D. (2005). Selection of K in K-means clustering. *Journal of Mechanical Engineering Science*, 205(1):103–119.

Solé, M., Muntés-Mulero, V., and Nin, J. (2012). Efficient microaggregation techniques for large numerical data volumes. *International Journal of Information Security - Special Issue: Supervisory control and data acquisition*, 11(4):253–267.

Zhou, C., Frankowski, D., Ludford, P., Shekhar, S., and Terveen, L. (2004). Discovering personal gazetteers: an interactive clustering approach. In *Proceedings of the annual ACM international workshop on Geographic information systems*, pages 266–273, New York, NY, USA.

A Cloud-based Exploration of Open Data: Promoting Transparency and Accountability of the Federal Government of Australia

Edwin Salvador
Department of Computing and Information Systems
University of Melbourne
Melbourne, Australia
edwin.salvador@epn.edu.au

Richard Sinnott
Department of Computing and Information Systems
University of Melbourne
Melbourne, Australia
rsinnott@unimelb.edu.au

Abstract

The Open Data movement has become more popular since governments such as USA, UK, Australia and New Zealand decided to open up much of their public information. Data is open if anyone is free to use, reuse and redistribute it. The main benefits that a government can obtain from Open Data include transparency, participation and collaboration. The aim of this research is to promote transparency and accountability of the Federal Government of Australia by using Cloud-related technologies to transform a set of publicly available data into human-friendly visualizations in order to facilitate its analysis. The datasets include details of politicians, parties, political opinions and government contracts among others. This paper describes the stages involved in transforming an extensive and diverse collection of data to support effective visualization that helps to highlight patterns in the datasets that would otherwise be difficult or impossible to identify.

1 Introduction

In recent years, the Open Data movement has become increasingly popular since the governments of various countries such as USA, UK, Australia, New Zealand, Ghana amongst many others decided to open up (some of) their data sets. In order to consider data as open, it should ideally be available preferably online in formats that are easy to read by computers and anyone must be allowed to use, reuse and redistribute it without any restriction (Dietrich et al., 2012). Furthermore, in the Open Data Handbook (Dietrich et al., 2012), the authors state that most of the data generated by governments are public data by law, and therefore they should be made available for others to use where privacy of citizens and national security issues are not challenged. According to the Open Government Data definition ("Welcome to Open Government Data," 2014), there are three main benefits that governments can obtain by opening up their data: transparency, participation and collaboration.

Acquiring and processing the amount of data generated by Governments may lead to workloads that are beyond the capacity of a single computer. Fortunately, the emergence of new technologies, such as Cloud Computing, makes it easier to scale the data processing demands in a seamless and scalable manner (Buyya, Yeo, Venugopal, Broberg, & Brandic, 2009). Whilst for some disciplines and domains where finer grained security is an impediment to adoption of Cloud computing, e.g. medicine, open data has by its very nature, no such impediments. Cloud computing also encourages the creation of more innovative services including those based on processing and analyzing datasets made available by governments. The sharing of technologies as open source solutions also goes hand in hand with open data initiatives.

The aim of this paper is to describe an approach taken to leverage the benefits provided by Open Data from the Australian government using Cloud-related technologies through the Australian national cloud facility: National eResearch Collaboration Tools and Resources (NeCTAR – www.nectar.org.au) and specifically the NeCTAR Research Cloud. The paper begins with a brief introduction to Open Data, providing its definition, its benefits and also its potential disadvantages. We then describe the advantages of using Cloud Computing to deal with Open Data. The details of the approach taken to harvest, clean and store publicly available data from Australian government resources followed by their analyses and visualizations of these datasets is given. Finally, the paper concludes by pointing out the importance of Open Government Data and the role of Cloud Computing to leverage the benefits offered by Open Data. It is emphasized that there is no causality implied in this paper regarding the analysis of the data offered. However we strongly believe that open discussions about causality are an essential element in the transparency of Government more generally.

2 Open Data

2.1 Definition

The Open Knowledge Foundation defines Open Data as ‘any data that can be freely used, reused and redistributed by anyone – subject only, at most, to the requirement of attribute and/or share-alike’ (Doctorow et al., 2014). We emphasize two important conditions that are not clearly mentioned in this short definition. First, data can be considered as open if it is easily accessible which means that data should be available on the Internet and in formats that are machine readable. Second, the terms reuse and redistribute include the possibility of intermixing two or more datasets in order to discover relations that would not be visible when having the datasets separated. The full definition provided by the Open Knowledge Foundation (Doctorow et al., 2014) gives further details of the conditions that should be satisfied by data to be considered as open. The final purpose of all these conditions required by Open Data is to ensure the potential interoperability of datasets, i.e. it is possible to combine any number of these datasets and subsequently identify their inter-relationships. Ideally this should be part of a larger system as opposed say to having many individual data sets (e.g. spreadsheets). The true power of Open Data is derived from the analytical tools and capabilities used to identify patterns that would otherwise remain hidden across multiple, diverse data sets.

2.2 Open Government Data

Governments are constantly gathering data from many types of sources: the population, taxes, quality of life indicators and indeed anything that could help the government to monitor and improve the management and governance of their country. Historically, only governmental entities (departments) have had access to process and analyze these data. However, according to (Davies, 2010; Dietrich et al., 2012; Lathrop & Ruma, 2010; Robinson, Yu, Zeller, & Felten, 2008), most of the data collected by government is public by law and therefore, it should be made open and available for everyone to use. In some cases, when governments have failed to make data easily accessible, citizens have had to find alternative ways to harvest and process these data to give it a meaningful use. A well-known case is the portal GovTrack.us which was launched in 2004 by a student who harvested a set of government data and published it in more accessible formats. This kind of initiatives have influenced in governments’ decisions to make government data publicly available (Brito, 2007; Hogge, 2010; Lathrop & Ruma, 2010). It should be noted also that government does not always share data effectively across its own departments – here the data includes both open and non-open data. The government departments of immigration, employment, education, health, transport, etc. all have subsets of the total “government” data, but the use of this data in integrated frameworks by government is currently lacking.

Since 2009, various countries have started Open Data initiatives by launching portals in which they publish government datasets to be downloaded by anyone. Among these countries are the USA (data.gov), the UK (data.gov.uk), Australia (data.gov.au), Ghana (data.gov.gh) and New Zealand (data.govt.nz). These sources of data are useful but do not include the tools to compare all of the data sets in any meaningful manner. Instead they are typically large collections of documents and individual (distinct) data sets. Often they are available as spreadsheets, CSV files with no means for direct comparison or analysis across the data sets.

2.3 Benefits

Many authors (Brito, 2007; Davies, 2010; Dietrich et al., 2012; Hogge, 2010; Lathrop & Ruma, 2010; Robinson et al., 2008) agree about the benefits that can be obtained by governments when they decide to open up their data, namely: transparency, participation and collaboration. These benefits are directly derived from the corresponding Open Data requirements: freedom of use, reuse and redistribution. In this context, the fact that anyone is free to use government data leads to an increment in government transparency. Hogge (2010), in her study mentions that transparency is not only about citizens trying to find irregularities in government actions, it is also about citizens constantly monitoring their governments' activities and providing feedback to improve processes and public services, and according to the Open Government Data definition ("Welcome to Open Government Data," 2014), this is what defines a well-functioning democratic society.

Open Data not only requires data to be accessible, but it requires the freedom to reuse these data for different purposes. This allows citizens to combine two or more datasets to create mash-ups and highlight potentially hidden relations between different datasets (Brito, 2007; Davies, 2010; Lathrop & Ruma, 2010). This improves the participation of citizens from different fields such as developers, scientists and indeed journalists. This is particularly important to governments since citizens can experiment in the creation of new services based on government data and the government is subsequently able to evaluate the most useful services and where appropriate shape future policy based on new knowledge. This has the added value of encouraging the participation of more citizens in government activities and increases the number of new services that could benefit the government.

The third key benefit of Open Data is collaboration which is directly derived from the freedom of users to redistribute government data, e.g. combining two or more datasets for a specific purpose and making the resulting dataset available for others to use. In this way, citizens are collaborating with each other while they are contributing to the government by creating services and solving problems. In some cases, this model of combining data sets to develop new, targeted solutions has spurned a range of start-ups and industries, e.g. San Francisco and the Civic Innovation activities (http://innovatesf.com/category/open-data/)

Although the process of making data publicly available can be seen as laborious and cost intensive to the government agencies involved, it brings further economic benefits to governments since it will improve the participation of people in the creation of innovative services (Hogge, 2010).

2.4 Barriers

According to (Davies, 2010; Lathrop & Ruma, 2010), transparency should not be focused only on the accountability and transparency of government. In fact, this could generate an excessive attention to government's mistakes and consequently, create an image of government as corrupt. This is clearly a reason why governments might not want to open up their data. However, the authors state that instead of reducing transparency, this problem could be addressed by creating a culture of transparency that not only judges when public entities behave badly, but a culture that is also capable to register approval when governments successfully solve public problems or deliver services in a cost effective manner.

Furthermore, many governments and indeed individuals are concerned about the privacy of citizens. Although, it is possible to anonymize datasets before they are made publicly available, it requires considerable time, effort and expense of public workers and sometimes it is not possible to guarantee that the data will be fully anonymized (Lathrop & Ruma, 2010). For this reason, some governments prefer to keep the data private. However it is the case that often terms such as protecting national security or citizen privacy are used as a blanket to deny access to many other data sets that are not contentious.

Additional barriers that stop governments making data publicly available is the fact that many data sets are stored on older forms of data storage media such as paper files and proprietary databases which do not allow for easy extraction and publication. Furthermore open data also requires appropriate (rich) metadata to describe it: the context in which it was collected, by whom and when. In some cases, this additional information is not directly available.

2.5 Disadvantages

Data can be open to misinterpretation, which can subsequently generate civic confusion and extra problems for governments. For instance, (Lathrop & Ruma, 2010) mentions a case where people correlated locations of crimes in a city with the number of immigrants in that location and make conclusions like "This is a high crime neighborhood because many immigrants live here”. Something which is not necessarily true as many other aspects must be taken into account to determine the reasons of high levels of crimes in a location.

Another disadvantage of publicly available data is for the potential for it to be manipulated with the intention of satisfying personal interests. This is difficult for a government to control and could be problematic since people often do not always verify data before making conclusions. Key to tackling this is the spirit of open data: it should be possible to verify or refute the conclusions that are drawn by access to the original data sets. Thus for any data that is accessed (harvested) it should always be possible to go back to the original (definitive) sources of the data (since it is open).

3 Cloud Computing

Open data benefits greatly by access to open data processing platforms. Cloud computing offers one approach that is directly suited to the processing of open data. The National Institute of Standards and Technology (NIST) (Mell & Grance, 2011), points out five essential characteristics that define the Cloud Computing model: on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service. In order to adapt to different types of users, Cloud providers offer three levels of abstraction: Software as a Service (SaaS) with examples being Salesforce’s CRM and Google Docs; Platform as a Service (PaaS) with examples being Microsoft Azure and Google App Engine, and Infrastructure as a Service (IaaS) with examples being Amazon EC2, Rackspace and the Australian NeCTAR Research Cloud. There are also many different Cloud deployment models: Private Clouds, Public Clouds, Community Clouds, and Hybrid Clouds (Mell & Grance, 2011; Sriram & Khajeh-Hosseini, 2010; Velte, Velte, & Elsenpeter, 2009; Zhang, Cheng, & Boutaba, 2010). Ideally open data should be processed on open Clouds and the applications and interpretation of the data utilizing open sources data models for complete transparency of the data and the associated data processing pipelines.

One of the main reasons for the success of Cloud computing is the capacity to rapidly scale up or scale down on demand, at an affordable cost and ideally in an automated fashion. This is particularly important when working with government data as they can become quite voluminous, they can change over time, they require veracity of information to be checks, and when comparisons and analyses are made between data sets these can result in computationally expensive requirements. Cloud Computing is especially suited to this environment since it is possible to scale out resources to satisfy needs and (in principle) pay for those extra resources only for the time that are actually being used. This is convenient specially for people considered ‘civil hackers’ who create services based on government data and most often without financial reward (Davies, 2010; Hogge, 2010). This contributes to the emergence of new questions and reduces the time needed to answer these questions, which encourages people to collect more data and create more innovative services.

The Cloud provider utilized here is the NeCTAR Research Cloud, which is an Australian government funded project that offers an IaaS platform with free access to Australian academics, or more precisely members of organizations subscribed to the Australian Access Federation (AAF – www.aaf.edu.au) such as the University of Melbourne. This work utilised two virtual machines (VMs) each with 2 cores, 8GB RAM and 100GB of storage. While the VMs were located in different zones, both have the same architecture (Figure 1). This allowed them to act as master at any time providing high availability to the system.

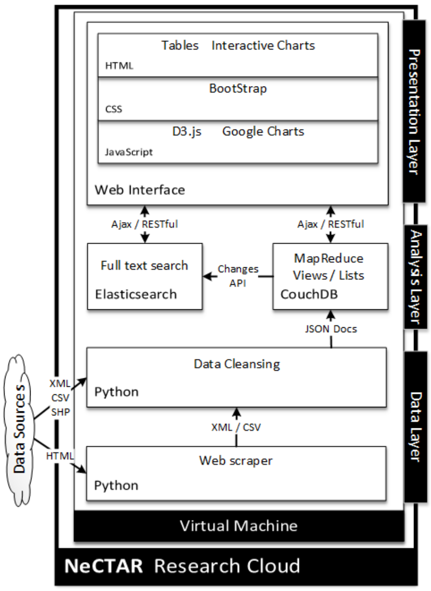


Figure 1. Architecture.

4 Implementation of the Open Government Data Access, Process and Visualise Platform

4.1 Data Layer

The key focus of the work is on access to and use of open government data. A *Data Layer* that harvested and processed these data was key to this. This layer was responsible for dealing with raw data coming from external sources. The data sets that were harvested and used as the basis for the work described here included:

- Australian Electoral Commission (www.aec.gov.au)
    - Annual Returns (2003 - 2013) (includes: party returns, political donations, Associated Entities, Political Expenditure)
    - Election Returns (2003 - 2013) (includes: donations to candidates, donors details, senate groups)
    - Election Results (2004 - 2010) (includes: Details of Federal election results divided in general, house and senate)
    - Federal electoral boundary GIS data (Census 2011)
- Portal data.gov.au
    - Historical Australian Government Contract Data (1999 - 2013)
    - Members of Parliament webpages and social networks
    - Portfolio of Responsibilities
- Parliament of Australia (www.aph.gov.au/Parliamentary_Business/Hansard)
    - House Hansard
    - Senate Hansard
- Lobbyists Details
    - Australian Government (www.lobbyists.pmc.gov.au)
    - Victoria (www.lobbyistsregister.vic.gov.au)
    - Queensland (www.lobbyists.integrity.qld.gov.au)
    - Western Australia (www.lobbyists.wa.gov.au)
    - Tasmania (www.lobbyists.dpac.tas.gov.au)
    - New South Wales (www.dpc.nsw.gov.au)
    - South Australia (www.dpc.sa.gov.au)

The analyses and visualizations of these data that drove and shaped the work were based on: political donations, election results, historical contracts data and political speeches. These data were selected following with researchers at the Centre for Advanced Data Journalism at the University of Melbourne. The Data Layer itself was divided into three stages: data harvesting, data cleansing and data storage which are described here.

Data Harvesting

It should be noted that most of the datasets that were harvested satisfy the requirements of Open Data, i.e. they are downloadable and are provided in machine-readable formats such as CSV and XML. It is also noted that there are other data that do not satisfy all of these requirements. For instance, the lobbyist registers for all the Australian States are available only in HTML (via web pages). In this case, it was necessary to implement web scrapers for webpages to extract the data and then store it in the database. This technique is inefficient and has several disadvantages for how data can be released as open data and subsequently used and interpreted. Firstly, it is error prone because a scraper may assume that a webpage follows a standard but there is the possibility of mistakes in the scraped HTML, which would cause the scraper to obtain erroneous data. Furthermore, it is a tedious task since it is almost impossible to build a scraper that works with many webpages as different sites use completely different designs. Lastly, the design of a webpage can change without any notice, which would render a given scraper totally useless and require a new scraper to be produced. Nevertheless, it is an effective technique when used carefully and after ensuring that all data obtained is verified before performing further analyses and interpretations. The information should also include metadata on when the data was accessed and scraped.

Additionally, even when data is made available in a more accessible (downloadable) format, further work is often required. For example, despite the fact that the Hansard political speeches of Australia are provided as downloadable XML files, there is no way to download the whole collection of speeches or the possibility of selecting a range of speech dates that could be downloaded. Consequently, it is often necessary to download one file at a time, which makes the process inefficient taking into account that there are thousands of files. As a result, whilst the data is open, the way in which it is made available is not really conducive to further processing without computational approaches to overcome these limitations, e.g. implementing processes to download all of the XML files.

It is recognized that the difficulties faced in harvesting public data are understandable since many governments (including the Australian government) are still in the process of opening their datasets and learning about how best to do this. These lessons are often spotlighted through important initiatives such as organized events used to receive feedback from data enthusiasts about how to improve the available datasets or which new datasets could be made available. For instance, GovHack is an annual event which brings together people from government, industry, academia and general public to experiment with government data and encourage open government and open data in Australia. Additionally, there exist various open data portals around Australia including the national portal data.gov.au, portals for every State such as the http://data.nsw.gov.au and even some cities like Melbourne have launched their own open data portals, e.g. https://data.melbourne.vic.gov.au/.

Data Cleansing

Every dataset collected will typically contain some extra and/or useless data that needs to be removed in order to improve the quality of data and increase the consistency between different datasets allowing them to be combined and interpreted more easily. To aid in data consistency, datasets from different formats such as CSV or XML were converted to JavaScript Object Notation (JSON) objects. Although, this process was simple, there were some difficulties to overcome in specific datasets. For instance, the XML files of the Hansard political speeches have different structures over different time periods, which made the process of parsing the whole collection more complex. However, it was possible to find certain levels of consistency in most of the datasets, which allowed use of Python scripts to convert hundreds of datasets and then store them in the database.

Data storage

Due to the variety of sources and the lack of a well-defined schema, CouchDB was selected as an appropriate database to store all the harvested data. CouchDB is a schema-free NoSQL and document-oriented database (Anderson, Lehnardt, & Slater, 2010). It stores its documents as JSON objects. In this model, each row of each dataset was stored as an independent JSON document adding an extra field “type”, and in some cases “subtype”, in order to facilitate the exploration of different datasets in the database.

Although both VMs were set up to act as a master at any given time, in this stage one VM could be considered as master and the other one as slave because only one of them could harvest data from external sources at a time while it replicated all the new data to the other. CouchDB provides strong replication processes that allow setting up a bi-directional replication between the databases in each VM. This allowed having both databases up to date while only one of them was harvesting data.

4.2 Analysis Layer

In addition to the flexibility provided by CouchDB to store schema-free documents, one of the main reasons to choose this database was its support for MapReduce based views. MapReduce is one of most effective approaches to deal with large-scale data problems and allows to separate what computations are performed and how those computations are performed (Buyya et al., 2009; Dean & Ghemawat, 2008; Ekanayake, Pallickara, & Fox, 2008; Lin & Dyer, 2010; Segaran & Hammerbacher, 2009; White, 2012). Therefore, to analyze the data the developer only needs to focus on the first part which consists on writing two functions: a map function and a reduce function. The run-time system handles how those computations are performed by managing failures, schedules and intercommunication. The complexity of map and reduce functions can be diverse and depends on the type of analysis to be performed on the data.

Furthermore, CouchDB documents and views are indexed using a B-Trees data structures, which are very efficient for storing large amounts of data (Anderson et al., 2010; Bayer, 1997). The index for a view is created only the first time that the view is queried and allows to retrieve large amount of data very quickly. In order reflect the current state of the database, the index of a view only needs to introduce the documents that have changed. Although this process is very efficient, it can introduce high latency to queries when a large amount of documents have changed (Anderson et al., 2010). This is a common problem faced by applications where documents in the database tend to be updated frequently. However, since the type of data used in this project is largely historical and not changing dynamically, CouchDB views were used successfully.

Most of the data analyses where performed using CouchDB views, these analyses included political donations over time, data aggregation of donations such as retrieving the largest donation in certain election period and correlation between different datasets, for instance, donations *vs* votes received by a political party. However, there were some cases where it was not possible to perform more complex analyses using only CouchDB views. For example, despite the fact that CouchDB owes many of its advantages to B-Trees, it also inherits one of its main drawbacks which is the inability to perform multi-dimensional queries (Bayer, 1997). In other words, CouchDB views are excellent to process queries such as the sum of donations received in the year 2004 (point queries) or the sum of donations received between 2004 and 2010 (range queries). However, for multi-dimensional queries such as the sum of donations received by a candidate from a specific donor in 2004 (4-dimensional), there were challenges that required support for other data processing capabilities. For this kind of query it was required to provide a visualization that showed an overview of the political donations. This visualization was required to group, color and filter donations in multiple ways and shows a summary for every group of donations. The summary includes the total sum of donations, the number of donations in that group, details of the largest donation and the top 3 donations received by candidates, parties and States. In order to solve this multi-dimensional query limitation, CouchDB functionalities were extended with ElasticSearch.

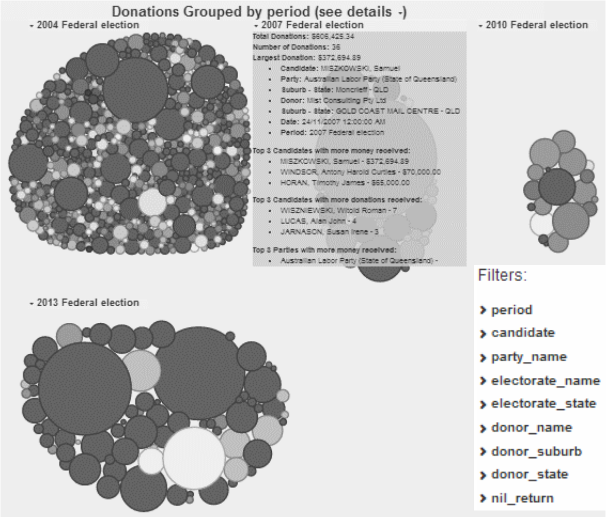
ElasticSearch is a distributed search engine built on top of Apache Lucene, which among other features provides full text search capabilities whilst hiding the complexity of Lucene behind a simple and coherent API. In spite of the document storage capabilities of ElasticSearch, it is mainly used as an extension for NoSQL databases thanks to a range of available plugins. For instance, it offers the possibility of indexing any CouchDB database through a plugin that listens to the changes API of CouchDB making the database searchable and allowing to perform more complex queries and more complete analyses of the data (Gormley & Tong, 2014). Using CouchDB in conjunction with ElasticSearch allows taking advantage of the most important features provided by each technology, namely durability and advanced search capabilities respectively. The number of features offered by ElasticSearch is vast and more details can be found in (Gormley & Tong, 2014).

ElasticSearch was also useful to build visualizations that correlate the political donations and the government contracts by searching all the occurrences of the donors' names in the dataset of government contracts. This was done through the Python API client provided by ElasticSearch and a Python script which returned a list of donor names that appeared in the government contracts dataset indicating the field where it was found, this helped to show the correlation of both datasets in a timeline.

4.3 Presentation Layer

Visualisation is essential when dealing with large-scale heterogeneous data sets. Indeed all of the data analyses would have limited value if it were not possible to visualize them in a human-friendly way. This is especially important in open government data initiatives where the focus is less on the detailed scientific model of discovery and more on the big picture questions that can be illustrated through the data itself. The presentation layer was based mainly in JavaScript using the D3.js library, Google Charts API and jQuery. In this section we illustrate some of the visualizations for the analyses mentioned previously.

Figure 2 shows an example of the multiple ways of visualizing political donations through one of the visualizations that were built. Each bubble in the figure represents a donation received by a candidate, the size of the bubble represents the amount of money donated, the color in this case, represents a political party and each group of bubbles is an election period. This is an interactive visualization so, donations could be grouped, colored and filtered by all the features contained in the dataset which include election period, candidate, party, electorate, electorate state, donor, donor state, donor suburb, and nil return. Furthermore, the labels for each group (including the main title) are clickable and they contain the summary for every group of donations and the main title contains the summary for all the four groups.



Total Donations: $606,425.34
Number of Donations: 36
Largest Donation: $372,694.89

- Candidate: MISZKOWSKI, Samuel
- Party: Australian Labor Party (State of Queensland)
- Suburb - State: Moncrieff - QLD
- Donor: Mist Consulting Pty Ltd
- Suburb - State: GOLD COAST MAIL CENTRE - QLD
- Date: 24/11/2007 12:00:00 AM
- Period: 2007 Federal election

Top 3 Candidates with more money received:

- MISZKOWSKI, Samuel - $372,694.89
- WINDSOR, Antony Harold Curties - $70,000.00
- HORAN, Timothy James - $65,000.00

Top 3 Candidates with more donations received:

- WISZNIEWSKI, Witold Roman - 7
- LUCAS, Alan John - 4
- JARNASON, Susan Irene - 3

Top 3 Parties with more money received:

- Australian Labor Party (State of Queensland) -

Figure 2. Overview of Donations.

This visualization facilitates a detailed exploration of the donations dataset and could be considered as a starting point for further analyses.

Another way of visualizing the donations is on a timeline as exposed in Figure 3. This shows the total number of donations received by date. Something interesting to point out here is how we can see that most of the peaks are in the years 2004, 2007 and 2010, years in which federal elections have taken place. This pattern of donations increasing in election years is also visible when looking at donations made by individual entities. Figure 4 illustrates all the donations made by an entity over time and highlights the tendency of making more donations in election years.

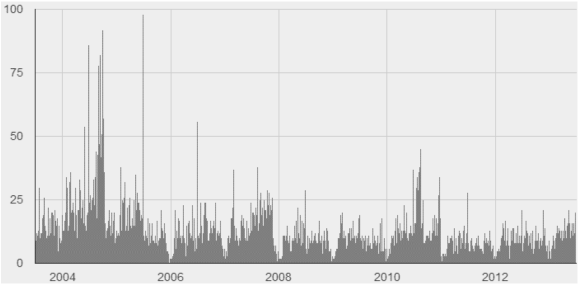
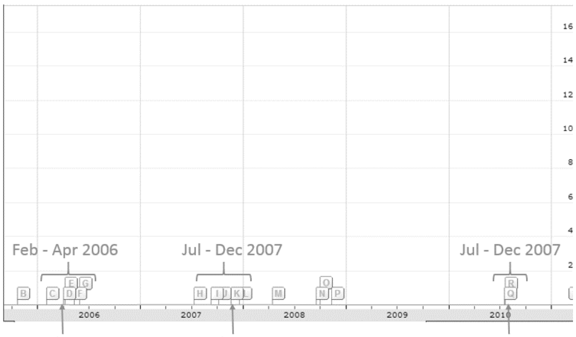


Figure 3. Summation of Political Donations Visualised over Timeline.



18th Feb: Elections SA-TAS

19th Nov: Federal Elections

21st Aug: Federal Elections

Figure 4. Individual Donations made by a given entity over time.

An additional scenario of interest is the correlation between political donations and government contracts, i.e. grants/award made to entities (most often companies). With the results obtained from the ElasticSearch analysis described in the previous section, donations and contracts were displayed in a timeline to explore whether the number of donations or the amount of money donated by an entity influenced (was correlated with) the number and the value of contracts that they subsequently obtained. Figure 5 shows this scenario for a specific entity.

It can be seen that there are several cases where contracts are obtained right before or after some donations have been made. In addition to the graph showed in Figure 5, this visualization also provides the details of the donations and contracts related with the entity being analyzed. Thus one can see the persons involved as well as political parties and governmental agencies and try to find more patterns to perform further investigations. For instance, a next step might be to investigate who is on the board of the companies making donations and to see if there exists a direct or indirect relation with the governmental agency that is offering the contract. It is emphasized that this is only one of the many scenarios that can be visualized with this analysis and there did not exist a clear correlation between the two datasets in many of the cases. However, this specific scenario helps us to demonstrate how mash-ups highlight hidden relations between apparently unrelated datasets. For transparency of government it is important to ensure that where major grants are awarded, independent review of political donations prior to the award can be scrutinized to ensure independence of government.

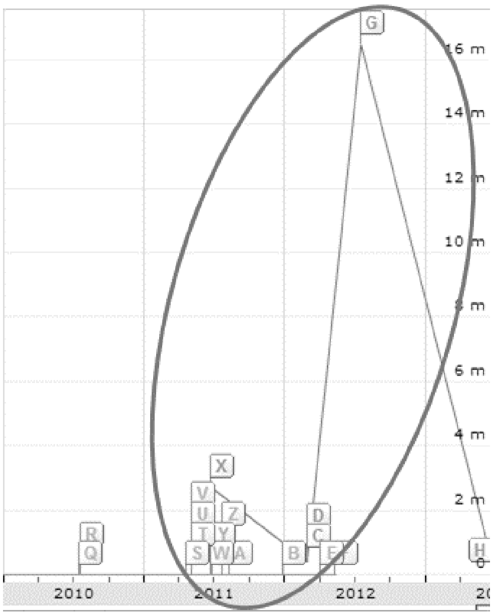


Figure 5. Colour-coded Correlation of Donations (A, B, E, F, Q-W, Y, Z) vs Government Contracts/Awards (C, D, G, H, X).

A further visualization is illustrated in Figure 6, which shows the correlation of terms used in political speeches over time. The figure demonstrate the correlation between the terms “boats” and “immigration” and it indicates how both terms tend to be used in the same dates. This visualization is useful to get an idea of what topics are being discussed by the members of the parliament in different periods.

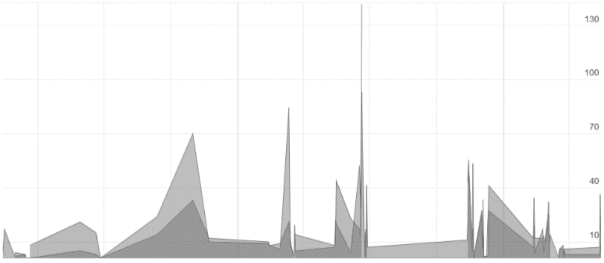


Figure 6. Political Speeches: correlation of words used over time.

An additional visualization using word clouds (Figure 7) was implemented to explore the most popular terms used by politicians in their speeches. This visualization allows to see a general word cloud for all the politicians in the collection of speeches and provides the possibility of filtering by year, month and day as well as selecting up to three politicians to show a comparison of the terms used by each of them over time. These word clouds provide a simple overview of the topics discussed by each politician. For instance, in Figure 7 the word cloud on the left belongs to the Shadow Minister for Transport and Infrastructure and so we can see that the most popular words are highly related to this charge such as “infrastructure”, “transport”, “highway”, “safety”, and “airport”. The word cloud on the right shows the words used by the Prime Minister in his speeches in May 2014 which is the month when the federal budget was presented to the parliament. In this case, we can see that the words “budget”, “deficit”, “spending”, and “tax” are amongst the most popular ones. This demonstrates that word clouds give us an idea of the topics that are dealt in parliament in different periods of time by different politicians. The combination of terms used in speeches and decisions made in award of contracts are also essential to correlate, e.g. speeches about the important of the Australian car industry should not be correlated/associated with political donations from car manufacturers for example if government is to be truly transparent and ultimately accountable for the independence of the decisions it makes.

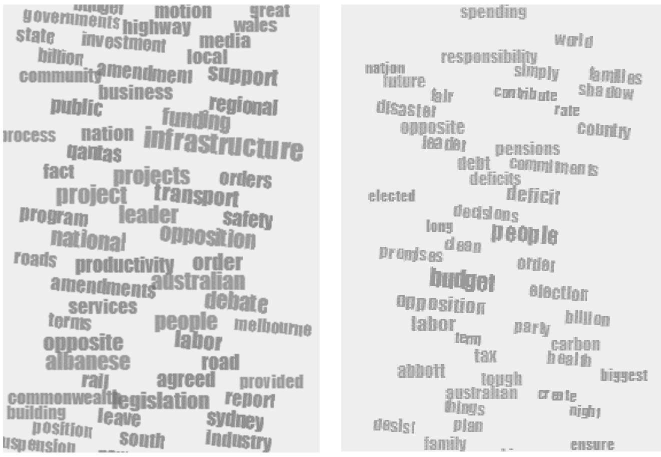


Figure 7. Word clouds comparing terms used by two politicians.

5 Conclusions

This paper presents an introduction to Open Data and points out how it could help governments to improve transparency and accountability. Furthermore, it describes some reasons why governments refuse to engage in Open Data initiatives as well as the existing disadvantages encountered if they are not managed correctly. The work described how and why Cloud Computing provide an appropriate environment for working with Open Data and identified and presented one of the many approaches that can be taken to set up this environment and the associated technologies involved. It also identified some of the common challenges faced by projects that deal with publicly available data and the methods used to overcome these challenges. Moreover, it showed multiple ways of visualizing data and how different datasets could be correlated to explore a portfolio of government data that is openly available on the web.

This work has many refinements that are currently ongoing. Incorporation of further data, e.g. membership of companies by politicians/their families/associates, as well as exploring social media use. The use of Twitter in particular offers a rich source of Open Data that can be accessed and used to help promote the overall information of government. Who is following whom on Twitter; who tweets on what topics; what is their sentiment on particular topics and how does this change over time are all on-going activities that are being pursued.

In all of this, it is emphasized that the purpose of this work is not to draw conclusions on any given government activity – this is the responsibility of others, e.g. investigative journalists. However for truly democratic and transparent governments it is essential that the data can be reviewed and analysed and stand up to public scrutiny. We strongly encourage this endeavor. All of the software and systems used in this work are also available. The existing prototype system is available at http://130.56.249.15/proj/.

Acknowledgments

The authors are thankful to the feedback and discussions that have shaped this work. Specifically we would like to thank Prof. Margaret Simons (Director of the Centre for Advanced Journalism at the University of Melbourne).

References

Anderson, J. C., Lehnardt, J., & Slater, N. (2010). *CouchDB: the definitive guide*: O'Reilly Media, Inc.

Bayer, R. (1997). The universal B-tree for multidimensional indexing: General concepts *Worldwide Computing and Its Applications* (pp. 198-209): Springer.

Brito, J. (2007). Hack, mash & peer: Crowdsourcing government transparency.

Buyya, R., Yeo, C. S., Venugopal, S., Broberg, J., & Brandic, I. (2009). Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. *Future Generation computer systems, 25*(6), 599-616.

Davies, T. (2010). Open data, democracy and public sector reform. *A look at open government data use from data. gov. uk. Über: http://practicalparticipation.co.uk/odi/report/wp-content/uploads/2010/08/How-is-open-governmentdata-being-used-in-practice. pdf.*

Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM, 51*(1), 107-113.

Dietrich, D., Gray, J., McNamara, T., Poikola, A., Pollock, R., Tait, J., et al. (2012). The Open Data Handbook. 2014, from http://opendatahandbook.org/en/

Doctorow, C., Suber, P., Hubbard, T., Murray-Rust, P., Walsh, J., Tsiavos, P., et al. (2014). The Open Definition. 2014, from http://opendefinition.org/

Ekanayake, J., Pallickara, S., & Fox, G. (2008). *Mapreduce for data intensive scientific analyses.* Paper presented at the eScience, 2008. eScience'08. IEEE Fourth International Conference on.

Gormley, C., & Tong, Z. (2014). *Elasticsearch: The Definitive Guide*: O'Reilly Media, Inc.

Hogge, B. (2010). Open data study. *a report commissioned by the Transparency and Accountability Initiative, available for download at: http://www.soros.org/initiatives/information/focus/communication/articles_publications/publications/open-data-study-20100519*.

Lathrop, D., & Ruma, L. (2010). *Open government: Collaboration, transparency, and participation in practice*: O'Reilly Media, Inc.

Lin, J., & Dyer, C. (2010). *Data-Intensive Text Processing with MapReduce*: Morgan and Claypool Publishers.

Mell, P., & Grance, T. (2011). The NIST Definition of Cloud Computing.

Robinson, D., Yu, H., Zeller, W. P., & Felten, E. W. (2008). Government data and the invisible hand. *Yale JL & Tech., 11*, 159.

Segaran, T., & Hammerbacher, J. (2009). *Beautiful data: the stories behind elegant data solutions*: O'Reilly Media, Inc.

Sriram, I., & Khajeh-Hosseini, A. (2010). Research agenda in cloud technologies. *arXiv preprint arXiv:1001.3259*.

Velte, T., Velte, A., & Elsenpeter, R. (2009). *Cloud computing, a practical approach*: McGraw-Hill, Inc.

Welcome to Open Government Data. (2014). 2014, from http://opengovernmentdata.org/

White, T. (2012). *Hadoop: The definitive guide*: " O'Reilly Media, Inc.".

Zhang, Q., Cheng, L., & Boutaba, R. (2010). Cloud computing: state-of-the-art and research challenges. *Journal of internet services and applications, 1*(1), 7-18.

Discovery of Sequential Patterns with Quantity Factors

Karim Guevara Puente de la Vega
Universidad Católica de Santa María /Arequipa
Universidad Nacional de San Agustín / Arequipa
kguevara@ucsm.edu.pe

Cesar Beltrán Castañón
Departamento de Ingeniería
Pontificia Universidad Católica del Perú / Lima
cbeltran@pucp.pe

Abstract

The sequential pattern mining stems from the need to obtain patterns that are repeated in multiple transactions in a database of sequences, which are related to time, or another type of criterion. This work presents the proposal of a new technique for the discovery of sequential patterns from a database of sequences, where the patterns not only provide information on how these relate to the time, but also, that in the mining process itself should be included the quantity factors associated with each of the items that are part of a sequence, and as a result of this process can be obtain information relating to how they relate these items with regard to the amounts associated. The proposed algorithm uses divide and conquer techniques, as well as indexing and partitioning of the database.

1 Credits

This document was written as part of the development of the 1st Symposium on Information Management and Big Data, SIMBig 2014. It has been adapted from the instructions for earlier ACL.

2 Introduction

The sequential pattern mining is the process by which you get the relationships between occurrences of sequential events, to find if there is a specific order in which these events occur. In relation to this area of study there are many investigations, all of them makes use of the restriction of minimal support, some include other restrictions, such as for example the time interval in which it is required that the events happen, also the use of taxonomies as defined by the user, and the fact of allowing the items in a sequence not necessarily must have occurred in a single transaction, but could be in two or more, always and when their times of each of these transactions is within some small window of time determined by the user.

In addition, the algorithms for mining sequential patterns of dealing with the previous sequential patterns in a uniform manner, despite the fact that these patterns individually in a sequence can have important differences such as the associated amount to each item that make up each pattern.

For the foregoing reasons, in the present paper proposes a technique by which it is intended to exploit these intrinsic relationships of the sequential patterns, in this specific case the relationship to the amount of each of the items. The inclusion of this aspect in the sequential pattern mining, you can afford to get a set of sequential patterns that are not only common but also let us know how these amounts associated with each item that is included in a sequential pattern frequent relates. The inclusion of the restriction of quantity within the extraction process of the frequent sequential patterns we could provide information much more meaningful.

The article is organized as follows: Section 2 is on the previous work. Section 3 gives a description of the problem. Section 4 introduces the technical proposal. Section 5 shows the experiments and results. The conclusions and future work are shown in section 6 and finally the references.

3 Previous works

The techniques of discovery of association rules are essentially boolean, due to which are discarded the quantities of the items purchased and only pay attention to if something was purchased or not. An important area of study is the sequential pattern mining that involves the extraction of patterns that are repeated in multiple transactions in a transactional database, which are related to time or another type of sequence.

The problem of the sequential pattern mining was introduced by Agrawal and Srikant (1995) set the example of the typical income of clients in a rental shop videos. Customers often rent "Star Wars", then "Empire Strikes Back" and then "Return of the Jedi". All these incomes not necessarily should have been made consecutively, that is to say, there could be customers that leased any other video in the middle of the previous sequence, so that these sequences of transactions also fall into the same pattern.

The researches on mining sequential patterns are based on events that took place in an orderly fashion at the time.

Most of the implemented algorithms for the extraction of frequent sequences, using three different types of approaches according to the form of evaluating the support of the candidate sequential patterns. The first group of algorithms is based on the ownership apriori, introduced by Agrawal and Srikant (1994) in the mining of association rules. This property suggests that any sub pattern from a frequent pattern is also frequent, allowing pruning sequences candidates during the process of lead generation. Based on this heuristics, Agrawal and Srikant (1995) proposed algorithms as the AprioriAll and AprioriSome. The substantial difference between these two algorithms is that the AprioriAll generates the candidates from all the large sequences found, but that might not be lowest panning values, however, AprioriSome only counts those sequences that are large but lowest panning values, thus reducing the search space of the patterns.

In subsequent work, Srikant and Agrawal (1996) propose the same algorithm GSP (Generalization Sequential Patterns), also based on the technical apriori, surpassing previous in 20 magnitudes of time. Until this time, the algorithms that had been proposed for mining sequential patterns focused on obtaining patterns taking into account only the minimal support given by the user. But these patterns could fit into transactions that had been given at intervals of time very distant, which was not convenient for the purposes of mining. So, in this paper, we propose the idea that in addition to the minimal support, the user could be in the ability to specify your interest in obtaining patterns that fit into transactions that have been given in certain periods of time, and this is made from the inclusion of restrictions on the maximum and minimum distance, the size of the window in which the sequences and the inheritance relationships - taxonomies, which are cross-relations through a hierarchy.

In these algorithms based on the principle of apriori, the greater effort focused on developing specific structures that allow sequential patterns represent the candidates and in this way make the counting operations support more quickly.

The second group is the algorithms that seek to reduce the size of the set of scanned data, by means of task execution of projection of the initial data base and the obtaining of patterns, without involving a process of lead generation. Using this technique and approach under the divide and rule, Han et al. (1996) proposed the algorithm FreeSpan (*Frecuent Pattern-Project Sequential Pattern mining*), and Pei et al. (2001) proposes PrefixSpan (*Prefix-projected Sequential Pattern mining*). In these algorithms the database of sequences is projected recursively in a set of small databases from which the fragments of sub sequences grow based on the current set of frequent sequences, where the patterns are extracted.

Han et al. (1996)] show that FreeSpan extracts the full set of patterns and is more efficient and considerably faster than the algorithm GSP. However, a sub sequence can be generated by the combinations of sub strings in a sequence, while the projection in FreeSpan must follow the sequence in the initial database without reducing the length. In addition, it is very expensive the fact that the growths of a sub sequence it will be explored in any point of the division within a candidate sequence. As an alternative to this problem, Pei (2001) proposes PrefixSpan. The general idea is to examine only the prefixes for the sub project only sequences and their corresponding sub sequences postfijas within databases planned. In each of these databases planned, it will find the sequential patterns expanded exploration only local patterns frequently. PrefixSpan extracts the full set of patterns and their efficiency and implementation are considerably better both GSP and FreeSpan.

The third group is formed by algorithms that kept in memory only information necessary for the evaluation of the bracket. These algorithms are based on the calls of occurrence lists that contain the description of the location where the patterns occur in the database. Under this approach, Zaki (2001) proposes the SPADE algorithm (*Sequential Pattern Discovery using Equivalence classes*) where he introduces the technical processing of the data base to vertical format, in addition there is a difference from the algorithms based on apriori, it does not perform multiple passes on the database, and you can extract all the frequent sequences in only three passes. This is due to the incorporation of new techniques and concepts such as the list of identifiers (id-list) with vertical format that is associated with the sequences. In these lists by means of temporary unions can be generated frequent sequences. Also used the grid based approach to break down the search space in small classes that can be processed independently in the main memory. Also, uses the search in both breadth and depth to find the frequent sequences within each class.

In addition to the techniques mentioned earlier, Lin and Lee (2005) proposes the first algorithm that implements the idea of indexing called Memisp memory (Memory Indexing for sequential pattern mining). The central idea of Memisp is to use the memory for both the data streams as to the indexes in the mining process and implement a strategy of indexing and search to find all frequent sequences from a sequence of data in memory, sequences that were read from the database in a first tour. Only requires a tour on the basis of data, at most, two for databases too large. Also avoids the generation of candidates and the projection of database, but presented as disadvantage a high CPU utilization and memory.

The fourth group of algorithms is composed of all those who use fuzzy techniques. One of the first work performed is the Wang et al. (1999), who propose a new data-mining algorithm, which takes the advantages of fuzzy sets theory, to enhance the capability of exploring interesting sequential patterns from the databases with quantitative values. The proposed algorithm integrates concepts of fuzzy sets and the AprioriAll algorithm to find interesting sequential patterns and fuzzy association rules from transaction data. The rules can thus predict what products and quantities will be bought next for a customer and can be used to provide some suggestions to appropriate supervisors.

Wang et al. (1999) propose fuzzy quantitative sequential patterns (FQSP) algorithm, where an item's quantity in the pattern is represented by a fuzzy term rather than a quantity interval. In their work an Apriori-like algorithm was developed to mine all FQSP, it suffers from the same weaknesses, including: (1) it may generate a huge set of candidate sequences and (2) it may require multiple scans of the database. Therefore, an Apriori-like algorithm often does not have a good performance when a sequence database is large and/or when the number of sequential patterns to be mined is large.

Chen et al. (2006) propose divide-and-conquer fuzzy sequential mining (DFSM) algorithm, to solve the same problem presented by Hong using the divide-and-conquer strategy, which possesses the same merits as the PrefixSpan algorithm; consequently, its performance is better than Wang et al.

Fiot (2008) in her work suggests that an item quantitative is partitioned into several fuzzy sets. In the context of fuzzy logic, a diffuse item is the association of a fuzzy set *b* to its corresponding item *x*, i.e. *[x,b]*. In the DB each record is associated with a diffuse item *[x,b]* according to their degree of membership. A set of diffuse items will be implicated by the pair *(X,B)*, where *X* is the set of items, and *B* is a set of fuzzy sets.

In addition, it argues that a sequence *g-k-sequence* ($s_1, s_2, \ldots, s_p$) is formed by *g* item sets diffuse *s=(X,B)* grouped to diffuse *k* items *[x,b]*, therefore the sequential pattern mining diffuse consists in finding the maximum frequency diffuse *g-k-sequence*.

Fiot (2008), provides a general definition of frequency of a sequence, and presents three algorithms to find the fuzzy sequential patterns: *SpeedyFuzzy*, which has all the objects or items of a fuzzy set, regardless of the degree, if it is greater than 0 objects have the same weight, *MiniFuzzy* is responsible for counting the objects or items of a fuzzy set, but supports only those items of the sequence that candidate have a greater degree of belonging to a specified threshold; and *TotallyFuzzy* that account each object and each sequence. In this algorithm takes into account the importance of the set or sequence of data, and is considered the best grade of membership.

4 Description of the Problem

A sequence *s*, denoted by $<e_1e_2\ldots i_n>$, is an ordered set of *n* elements, where each element e_i is a set of objects called *itemset*. An *itemset*, which is denoted by ($x_1 [c_1], x_2 [c_2], \ldots, X_q[c_q]$), is a non-empty set of elements *q*, where each element x_j is an item and is represented by a literal, and c_j is the amount associated with the item x_j that is represented by a number in square brackets. Without loss of generality, the objects of an element are supposed to be found in lexicographical order by the literal. The size of the sequence *s*, denoted by $|s|$, is the total number of objects of all elements of the *s*, so a sequence *s* is a *k-sequence*, if $|s|=k$.

For example, *<(a[5])(c[2])(a[1])>*, *<(a[2],c[4])(a[3])>* and *<(b[2])(a[2],e[3])>* are all 3-sequences. A sequence $s = <e_1e_2\ldots i_n>$ is a sub-sequence of another sequence of $s'=<e_1'e_2'\ldots e_m'>$ if there are $1\le i_1<i_2<\ldots<i_n\le m$ such that $e_1\subseteq e_{i1}'$, $e_2\subseteq e_{i2}'$, ... , and $e_n\subseteq e_{in}'$. The sequence *s'*

contains the sequence *s* if *s* is a sub-sequence of *s'*.

Similarly, *<(b,c)(c)(a,c,e)>* contains *<(b)(a,e)>* where the quantities may be different.

The support (*sup*) of a sequential pattern X is defined as the percentage on the fraction of records that contains X the total number of records in the database. The counter for each item is increased by one each time the item is found in different transactions in the database during the scanning process. This means that the counter of support does not take into account the quantity of the item. For example, in a transaction a customer buys three bottles of beer, but only increases the number of the counter to support {beer} by one; in other words, if a transaction contains an item, then, the support counter that item only is incremented by one.

Each sequence in the database is known as a sequence of data. The support of the sequence *s*, is denoted as *s.sup*, and represents the number of sequences of data that contain *s* divided by the total number of sequences that there is in the database. *minSup* threshold is the minimum specified by the user. A sequence *s* is frequent if *s.sup≥minSup*, therefore it will be a sequential pattern frequently.

Then, given the value of the *minSup* and a database of sequences, the problem of the sequential pattern mining is to discover the set of all sequential patterns whose supports are greater equal to the value of the minimum support (*s.sup≥ minSup*).

Definition: given a *ρ* pattern and a frequent item *x* in the database of sequences, *ρ'* is a:

- **Pattern Type-1**: if *ρ'* can be formed by adding to *ρ* the itemset that contains the item *x*, as a new element of *ρ*.
- **Pattern Type-2**: if *ρ'* can be formed by the extension of the last element of *ρ* with *x*.

The item x is called *stem* of the sequential pattern *ρ'*, and *ρ* prefix is the pattern of *ρ'*.

That is, the following database of sequences of figure 1, which includes amounts for the items and that, has six sequences of data.

Sequences
C1 = <(a[1],d[2]) (b[3],c[4]) (a[3],e[2])>
C2 = <(d[2],g[1]) (c[5],f[3]) (b[2],d[1])>
C3 = <(a[5],c[3]) (d[2]) (f[2]) (b[3])>
C4 = <(a[4],b[2],c[3],d[1]) (a[3]) (b[4])>
C5 = <(b[3],c[2],d[1]) (a[3],c[2],e[2]) (a[4])>
C6 = <(b[4],c[3]) (c[2]) (a[1],c[2],e[3])>

Figure 1. Database of sequences

Consider the sequence *C6*, which consists of three elements, the first has the objects *b* and *c*, the second has the object *c*, and the third has the objects *a, c,* and *e*. Therefore, the support of *<(b)(a)>* is 4/6 since all the sequences of data with the exception of *C2* and *C3* contain a *<(b)(a)>*. The sequence *<(a,d)(a)>* is a sub sequence of both *C1* and *C4*; and therefore, *<(a,d)(a)>.sup=2/6*.

Given the pattern *<(a)>* and the frequent item *b*, gets the pattern type-1 *<(a)(b)>* adding *(b)* to *<(a)>*, and the pattern type-2 *<(a,b)>* by the extension of *<(a)>* with *b*.

Similarly, *<(a)>* is the prefix pattern (*ρ_pat*) which in turn is a frequent sequence, and *b* is the stem of both: *<(a)(b)>* and *<(a,b)>*.

Note that the sequence *null*, denoted by *<>*, is the *ρ_pat* of any *1-frequent sequence*. Therefore, a *k-sequence* is like a frequent pattern type-1 or type-2 of a (*k-1)-frequent sequence*.

5 Algorithm for the discovery of sequential patterns with quantity factors - MSP-QF

The algorithm for mining sequential patterns with quantity factors, arises from the need to discover from a database of sequences, the set of sequential patterns that include the amounts associated with each of the items that are part of the transactions in the database, since having this additional information can be known with greater precision not only what is the relationship with respect to the time that exists between the various items involved in the transactions of a sequence, but also as is the relationship to the amount of these items.

The algorithm MSP-QF, it is based on the idea of the use of prefixes, and the creation of indexes from the database of sequences or other indices that are generated during the mining process, where recursively searching for frequent patterns. As a result of the exploration of a particular index, fewer and shorter sequences of data need to be processed, while the patterns that are found will be made longer.

In addition, if the database is very large sequence uses the techniques of partitioning in a manner that the algorithm is applied to each of the partitions as if it were a database of lesser size.

5.1 Procedure of the algorithm MSP-QF

Listed below are the steps of the proposed algorithm.

Step 1: Partitioning and scanning of the database of sequences. Depending on the size of the database are applicable to so it can be partitioned and formatted through and then to scan each of the partitions of independently. For each partition, the sequences are constructed and stored in the structure *DBSeq*. At the same time generates the index of items where is stored the support for each one of them, which is found during the scanning process.

Step 2: The index of items are filtered out those that are frequent, i.e., whose support is greater than or equal to minSup determined by the user. All these items come to form sequences of size |s| =1, therefore, form the set of 1-sequences. For all these sequences frequent item is to write the amounts associated with each item to the time it is saved in the whole of frequent patterns.

Step 3: For each one of the frequent patterns ρ, found in step 2, or as a result of the step 4, the index is constructed ρ_idx, with inputs (*ptr_ds*, *pos*), where *ptr_ds* refers to a sequence of the DB in which appears the ρ pattern, and *pos* is the pair (*posItemSet*, *posItem*), where *posItemSet* is the position of the *itemset* in the sequence and *posItem* the position of the item in the itemset from the sequence where the pattern appears. The values of *pos* allow the following scans are performed only on the basis of these positions in a certain sequence.

Step 4: Find the stems of type-1 and/or type-2 for each ρ pattern and its corresponding index ρ*_idx* generated in the previous step, considering only those items of the sequences referred to in ρ*_idx* and the respective values of pos. At the same time as are the stems are calculated their supports, and in addition is added to the list of quantities of the item that is part of the stem the amount referred to in the item of the sequence of the DB which is being examined. The information of the stems and their quantities are stored in another index of stems. This step is repeated for the same pattern, until they were no longer more stems from this.

Step 5: When there is no more stems, filtered index stems all those who are frequent. For all stems (sequences) frequently, we proceed to discretize the quantities that were associated with each item and stored in the set of frequent patterns. For this, before adding it to the set of frequent patterns, we proceed to verify that the common pattern found recently has not already been added before this set as a result of applying the algorithm to a partition of the database that was processed with previously. If frequent pattern already exists in the set of frequent patterns, the discretization process is again applied to the set of quantities associated with the sequence is stored as a frequent pattern and set of quantities of newly discovered frequent pattern; otherwise, the common pattern found in the current partition is added directly to the set of frequent patterns.

Then we proceed to perform recursively steps 3, 4 and 5 with each one of the frequent patterns that are found in the process.

Discretization Function: This function is responsible for making the set of quantities associated with an item, the range of values given by the mean and standard deviation of this set. For example, given the sequences of the figure 1, the set of quantities associated with the item <(a)> is: 1,5,4,3,1, which after being discretized would be the interval formed by: [2.8±1.6]

To summarize the steps carried out in the proposed algorithm, figure 3 shows a schematic of the entire procedure.

5.2 Algorithm specification MSP-QF

Here we show the specification of the proposed algorithm MSP-QF.

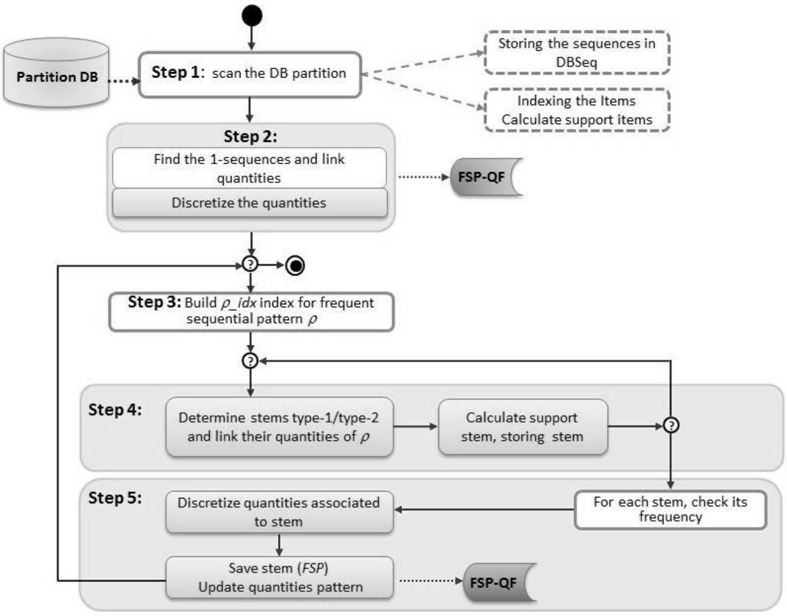
```
Algorithm MSP-QF
In:  DB = database sequences
     minSup = minimum support
     partition = number of sequences included in each of the partitions
Out: set of all sequential patterns with quantity factors.
Procedure:
1. Partitioning the DB
2. Each partition scan it in main memory and:
     (i) build sequences and store them in DBSeq structure.
     (ii) index the items and determine the support of each item.
     (iv) associate the quantities of each item in a sequence list of item
          quantities in the index.
3. Find the set of frequent items
4. For each frequent item x,
     (i) form the sequential pattern $\rho$ = <(x)>
     (ii) call Discretize($\rho$) to discretize the set of quantities associated
          with each item $x \in \rho$.
     (iii) storing $\rho$ in the set frequent patterns.
     (iv) call Indexing (x, <>, DBSeq) to build the $\rho$-idx  index.
     (v) call Mining ($\rho$, $\rho$-idx) to obtain patterns from index $\rho$-idx.
```

```
**Subrutine** *Indexing* (*x*, $\rho$, *set_Seq*)
**Parameters:**
    x = one stem type-1 or type-2;
    $\rho$ = prefix pattern ($\rho$-*pat*);
    *set_Seq* = set of data sequences
    / * If *set_Seq* is an index, then each data sequence in the index is
    referenced by the element *ptr_ds¸* which is formed at the input
    (*ptr_ds, pos*) index * /
**Out:** índex $\rho$'-*idx*, where $\rho$' represents the pattern formed by the stem *x*
and prefix pattern $\rho$-*pat*.
**Procedure:**
1. For each data sequence *ds* of *set_Seq*
    (i) If *set_Seq* = *DBSeq* the *pos_inicial* = *0*, else *pos_inicial* = *pos*.
    (ii) Find the stem in each sequence *ds* from the position
    (*pos_inicial* + *1*),
        1. If the stem *x* is in position *pos* in ds, then insert a pair (*ptr_ds,*
        *pos*) in $\rho$'-idx index, where *ptr_ds* reference to *ds*.
        2. If the stems *x* is equal to the item *x'* of the *ds* sequence, add-
        ed the quantity *q* associated with the item x', to the list of
        quantities related to x.
2. Return the $\rho$'-*idx* index.
```

```
**Subrutine** *Mining*($\rho$,$\rho$-*idx*)
**Parameters:**
    $\rho$ = a pattern;
    $\rho$-*idx* = an índex.
**Procedure:**
1. For each data sequence *ds* referenced by *ptr_ds* of input *(ptr_ds,*
    *pos)* in $\rho$-*idx*,
    (i) Starting from the (*pos +1*) position until |ds|, determining poten-
        tial stems and increase in one support each of these stems.
2. Filter those stems that have a large enough support.
3. For each stem x found in the previous step,
    (i) form a sequential pattern $\rho$'' from the prefix pattern $\rho$-*pat* and
        the stem *x*.
    (ii) call Discretize($\rho$') to discretize the amounts associated with
        the items of $\rho$'.
    (iii) call Index($\rho$', $\rho$, $\rho$'-idx) to build the index $\rho$'-idx.
    (iv) call Mining($\rho$', $\rho$'-idx) to discover sequential patterns from
        index $\rho$'-idx
```

```
**Subrutina** *Discretize*($\rho$)
**Parameters:**
    $\rho$ = a pattern that is a sequence;
**Output:** the arithmetic mean and standard deviation of the amounts
associated with each item $\rho$ pattern.
**Procedure:**
1. For each itemset $\gamma \in \rho$ do
    a) For each item $x \in \gamma$ do
        (i) Calculate the arithmetic mean and standard deviation of
            the set of quantities associated with the item *x*
        (ii) storing the arithmetic mean and standard deviation in the
            pattern $\rho$
```

Figure 2. Specification of the algorithm MSP-QF



Partition DB
Step 1: scan the DB partition
Storing the sequences in DBSeq
Indexing the Items
Calculate support items
Step 2:
Find the 1-sequences and link quantities
Discretize the quantities
FSP-QF
Step 3: Build ρ_idx index for frequent sequential pattern ρ
Step 4:
Determine stems type-1/type-2 and link their quantities of ρ
Calculate support stem, storing stem
Step 5:
Discretize quantities associated to stem
For each stem, check its frequency
Save stem (FSP)
Update quantities pattern
FSP-QF

Figure 3. Schema of the procedures of the algorithm MSP-QF

6 Experiments and Result

The experiments to test the technical proposal were implemented in two different scenarios, which are described below.

6.1 Scenario 1: Real data

The technique was applied in the analysis of the market basket of a supermarket. These tests consisted of obtain the set of frequent sequential patterns from the basis of data obtained in the course of three non-consecutive periods. The first period goes from mid-December of 1999 until mid-January 2000. The second period goes from early 2000 until the beginning of June of the same year. The third period goes from late August 2000 until the end of November 2000. This database consists of 88163 transactions, 3000 items unique to approximately 5133 customers.

The purpose of testing is to discover patterns of customer usage in the supermarket, plus get the amount of each of the items that will be purchased by these customers as a result of applying the proposed technique, which will allow us to have more accurate and significant in terms of the quantity purchased of each of the items.

Seven tests were carried out with minimum media 10 %, 2%, 1.5%, 1.0%, 0.75%, 0.50% and 0.25%, which were observed in figure 4. These results were compared with results of the technical Memisp.

minSup (%)	MEMISP		MSP-QF	
	Exe.Time (seg.)	No. Patterns	Exe.Time (seg.)	No. Patterns
10.00	4	50	5	50
2.00	12	824	15	824
1.50	16	1371	19	1371
1.00	22	2773	27	2773
0.75	28	4582	35	4582
0.50	39	9286	50	9286
0.25	72	30831	89	30831

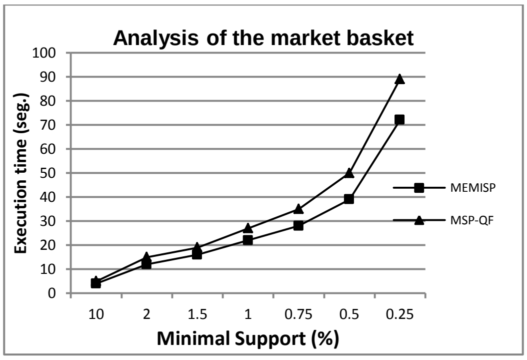


Figure 4. Results of the tests for scenario 1

In the test with *minSup=2%* were obtained 824 sequential patterns with quantity factors, some of which are:

```
Olive[ 1.06±0.51 ]$
Olive[ 1.03±0.5 ]$ Paprika[ 0.37±0.23 ]$
Olive[ 1.01±0.5 ]$ Porro[ 0.57±0.27 ]$
Celery[ 0.56±0.25 ]$ Lettuce[ 0.53±0.24 ], Lentils[ 0.54±0.23 ]$
Celery[ 0.56±0.26 ]$ Lettuce[ 0.55km Air ±0.24 ],
                                        Paprika[ 0.34±0.17 ]$
Lettuce[ 0.61±0.25 ], Lentils[ 0.56±0.26 ]$ Porro[ 0.59±0.24 ],
                                        Paprika[ 0.33±0.15 ]$
Porro[ 0.54±0.27 ], Lentils[at 0.62±0.25 ]$ Lentils[ 0.58±0.25 ]$
                                       Paprika[ 0.35±0.17 ]$
```

Of these sequential patterns we can clarify the following with regard to purchases made by customers:

- Customers buy only olives in a quantity of 1.06±0.51.
- Customers who have purchased a first time only olive, returning a next time for chili or by porro, with quantities of 0.37±0.23 and 0.57±0.27 respectively . Those who buy after pepper, purchased before olives in a quantity equal to 1.03±0.5, while those who acquire porro did so with an amount equal to 1.01±0.5.
- Those who buy lettuce at the same time buy lentils in amounts equal to 0.61±0.25 and 0.56±0.26 respectively. Later, these same customers buy porro and paprika with amounts equal to 0.59±0.24 and 0.33±0.15.
- Those who buy porro, in the same transaction also buy lentils. Later return to buy only lentils, and a next time buy only paprika, in the amounts listed in the pattern.

6.2 Scenario 2: Synthetic data

This second scenario is generated multiple databases (datasets) of synthetic form by means of the Synthetic Data Generator Tool.

The process followed to synthetic generation of the dataset, it is the describing Agrawal and Srikan (1995), and under the parameters referred to in the work of Lin and Lee (2005).

In this scenario, tests were carried out both of effectiveness, efficiency and scalability.

The evidence of effectiveness and efficiency were made with dataset generated with the following parameters: $NI = 25000$, $NS = 5000$, $N = 10000$, $|S| = 4$, $|I| = 1.25$, $corr_S = 0.25$, $crup_S = 0.75$, $corr_I = 0.25$ and $crup_I = 0.75$.

The results of these tests were compared with the results obtained for the algorithms PrefixSpan-1, PrefixSpan-2 and Memisp.

Efficiency Tests: Ran a first subset of tests for $|C|=10$ and a database of *200,000* sequences, with different values for *minSup*. The results are shown in figure 5.

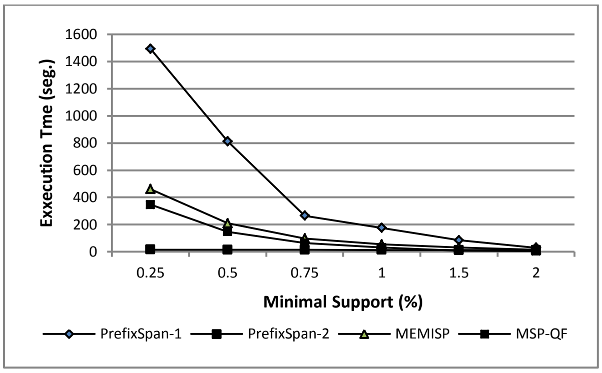


Figure 5. Test results for |C|=10 and |DB|=200K

The second subgroup of tests was conducted with a dataset with values for |C| and |T| of 20 and 5 respectively. This value of |T| implies that the number of items of transactions increases, which represents that the database is also larger and more dense with respect to the number of frequent sequential patterns that may be obtained. The results of these tests are those seen in figure 6.

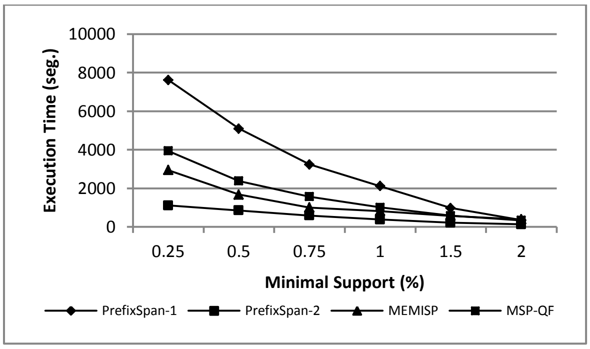


Figure 6. Test results for |C|=20, |T|=5 and |DB|=200K

A last subset of efficiency tests were carried out under the same parameters of the subset above with the exception of |T| increased to 7.5. The results are shown in figure 7.

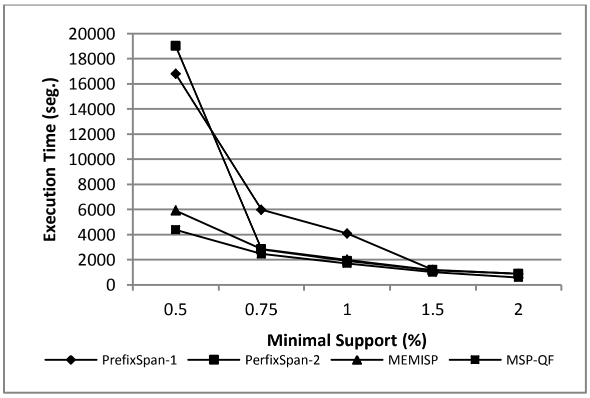


Figure 7. Test results for |C|=20, |T|=7.5 and |DB|=200K

Efficacy tests: Were carried out 92 efficacy trials with the same dataset of the tests of efficiency. In four of the tests carried out with values |C| =20 and |T| equal to 2.5, 5 and 7.5 respectively, it did not achieve the same amount of sequential patterns found with the algorithms PrefixSpan-1 and PrefixSpan-2. These 4 tests represent 4% of the total.

Scalability tests: The scalability tests were used datasets synthetically generated with the same values of the parameters of the first subset of tests of efficiency, and with minimal support equal to 0.75%. The amount of sequences in the dataset for these tests ranged from |DB|=1000K to 10000K, i.e. of a million to 10 million sequences. In figure 8, you can watch the results of these tests.

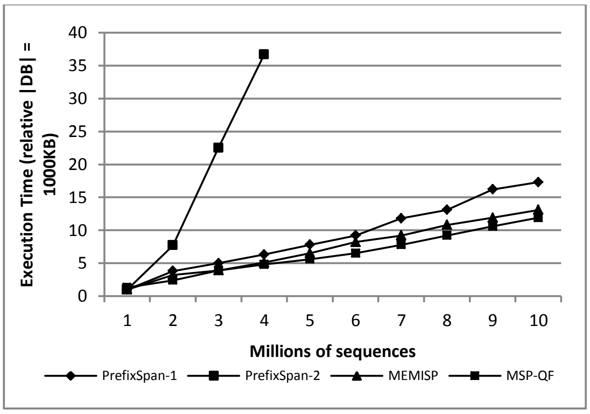


Figure 8, Results of scalability tests for *min-sup=0.75%* and different sizes of datasets

7 Conclusion

We have proposed an algorithm for the discovery of sequential patterns that allows, as part of the mining process, to infer from the amounts associated with each of the items of the transactions that make up a sequence, the quantity factors linked to the frequent sequential patterns.

The technical proposal has been designed in such a way that uses a compact set of indices in which focuses the search for the sequential patterns from frequent patterns that have already been found earlier and that represent the prefixes of the patterns to find. That is why the size of the indexes is decreasing in accordance with the mining process progresses.

In addition, there has been that the information provided by the frequent patterns with factors of quantity, is much more accurate, since not only gives us information on how is the temporal relationship of the items in the various transactions, but also, what is the relationship of the quantities of some items to others, which enriches the semantics provided by the set of sequential patterns.

Finally, the results obtained in section 5, we can conclude by saying that the technical proposal meets the objectives of the mining process; it is effective, is efficient and is scalable because it has a linear behavior in accordance with the sequence database grows, and that when applied to large data bases his performance turned out to be better than the techniques discussed in this work.

Reference

Agrawal Rakesh and Srikant Ramakrishnan. 1994. *Fast algorithms for mining association rules*. In Proceeding 20th International Conference Very Large Data Bases, VLDB.

Agrawal Rakesh and Srikant Ramakrishnan. 1995. *Minning Pattern Sequential*. 11th International Conference on Data Engineering (ICDE'95), Taipei, Taiwan.

Agrawal Rakesh and Srikant Ramakrishnan. 1996. *Mining Quantitative Association Rules in Large Relational Tables*. In Proceeding of the 1996 ACM SIGMOD Conference, Montreal, Québec, Canada.

Alatas Bilal, Akin Erhan and Karci Ali. 2008. *MODENAR: Multi-objective differential evolution algorithm for mining numeric association rules*. Applied Soft Computing.

Chen Yen-Liang and Haung T. Cheng-Kui. 2006. *A new approach for discovering fuzzy quantitative sequential patterns in sequence databases*. Fuzzy Sets and Systems 157(12):1641–1661.

Fiot Céline. 2008. *Fuzzy Sequential Patterns for Quantitative Data Mining*. In Galindo, J. (Ed.), Handbook of Research on Fuzzy Information Processing in Databases.

Han Jiawei, Pei Jian, Mortazavi-Asl Behzad, Chen Qiming, Dayal Umeshwar and Hsu Mei-Chun. 1996. *Freespan: Frequent pattern-projected sequential pattern mining*. Conference of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining.

Karel Filip. 2006. *Quantitative and Ordinal Association Rules Mining (QAR Mining)*. 10th International Conference on Knowledge-Based & Intelligent Information & Engineering Systems (KES 2006). South Coast, UK: Springer, Heidelberg.

Lin Ming-Yen and Lee Suh-Yin. 2005. *Fast Discovery of Sequential Patterns through Memory Indexing and Database Partitioning*. Journal of Information Science and Engineering.

Market-Basket Synthetic Data Generator, http://synthdatagen.codeplex.com/.

Molina L. Carlos. 2001. *Torturando los Datos hasta que Confiesen*. Departamento de Lenguajes y Sistemas Informáticos, Universidad Politécnica de Cataluña. Barcelona, España.

Papadimitriou Stergios and Mavroudi Seferina. 2005. *The fuzzy frequent pattern Tree*. In 9th WSEAS International Conference on Computers. Athens, Greece: World Scientific and Engineering Academy and Society.

Pei Jian, Han Jiawei, Mortazavi-Asl Behzad and Pinto Helen. 2001. *PrefixSpan: Mining sequential patterns efficiently by prefix-projected pattern growth*. In ICDE '01 Proceedings of the 17th International Conference on Data Engineering.

Srikant Ramakrishnan and Agrawal Rakesh. 1996. *Mining Sequential Patterns: Generalizations and Performance Improvements*. In Proc.5th Int. Conf. Extending Database Technology (EDBT'96), pages 3–17, Avignon, France.

Takashi Washio, Yuki Mitsunaga and Hiroshi Motoda. 2005. *Mining Quantitative Frequent Itemsets Using Adaptive Density-Based Subspace Clustering*. In Fifth IEEE International Conference on Data Mining (ICDM'05). Houston, Texas, USA: IEEE Computer Society.

Wang Shyue, Kuo Chun-Yn and Hong Tzung-Pei. 1999. *Mining fuzzy sequential patterns from quantitative data*", 1999 IEEE Internat. Conference Systems, Man, and Cybernetics, vol. 3, 1999, pp. 962–966.

Zaki Mohammed J. 2001. *SPADE: An efficient algorithm for mining frequent sequences*. Machine Learning, 42(1-2):31–60.

Online Courses Recommendation based on LDA

Rel Guzman Apaza, Elizabeth Vera Cervantes, Laura Cruz Quispe, José Ochoa Luna
National University of St. Agustin
Arequipa - Perú
{r.guzmanap,elizavvc,lvcruzq,eduardo.ol}@gmail.com

Abstract

In this paper we propose a course recommendation system based on historical grades of students in college. Our model will be able to recommend available courses in sites such as: Coursera, Udacity, Edx, etc. To do so, probabilistic topic models are used as follows. On one hand, Latent Dirichlet Allocation (LDA) topic model infers topics from content given in a college course syllabus. On the other hand, topics are also extracted from a massive online open course (MOOC) syllabus. These two sets of topics and grading information are matched using a content based recommendation system so as to recommend relevant online courses to students. Preliminary results show suitability of our approach.

1 Introduction

Nowadays, the amount of educational resources spread at Internet is huge and diverse (Martin, 2012). Massive Online Open Courses (MOOCs) such us Coursera, Udacity, EdX, to name a few, are gaining momentum (Fischer, 2014). It is possible to find courses from almost every knowledge domain. This vast offer overwhelm any user willing to find courses according his/her background. This task can be tedious because it involves access to each platform, search available courses, select some courses, read carefully each course syllabus, and choose appropriate content. This process can be unmanageable if we extend our search beyond online courses to educational content.

In this work we propose a system for online courses recommendation, although MOOCs courses are primarily focused. To do so, we rely on Topic Models (Blei, 2012), an unsupervised probabilistic generative model, which given a set of documents and a number of topics as input, automatically returns a relevant set of words probabilistically associated for each topic. Why this scheme is valuable?, consider for instance a huge number of digitalized books of a public library, this algorithm can automatically discover main topic words and therefore allows one to gain insights about content in books.

Currently educational systems and data mining is an emerging research area (Romero and Ventura, 2010), these systems use different recommendation techniques in order to suggest online learning activities, based on preferences, knowledge and data from other students with similar interests (Romero et al., 2007). In (Kuang et al., 2011) the author provides resource recommendation for users in the e-learning system based on contents and user log activities. There was proposed a method for resource recommendation based on topic modeling in an e-learning system, that system used Latent Dirichlet Allocation (LDA) to get a low dimension vector, and to do inference it used Gibbs sampling, then in resource recommendation it applied cosine similarity in document topic distribution to find neighbor resources. The authors from (Haruechaiyasak and Damrongrat, 2008) also recommended documents, in this case it recommended articles from wikipedia by calculating the similarity measures among topic distributions of the articles. The model proposed in (Sadikov and Bratko, 2011) is an hybrid recommendation system where the core of the system is a linear regression model, based on stochastic gradient descent. For predicting the rank of a lecture, they used and compared the predictions made by content-based and collaborative-based methods. In this paper they established manually the attributes that represent each video-lecture, unlike our paper, where the attributes for the courses are defined by the LDA algorithm. In (Sadikov and Bratko, 2011), to find a rank they measured the correlation between an old lecture (a lecture the visitor has already seen), and the new lectures (lectures that visitor has not seen yet), and then they ordered theses measures in a list, where the lowest comes first, theses computations were used in the linear regression model. Also they said that there was not to much difference between using content-based or collaborative-based methods, but they said that their system could have been improved if they used textual attributes, which is our case.

In our proposal, Latent Dirichlet Allocation (LDA) (Blei et al., 2003) topic model is mainly used as feature descriptor of courses. Thus, we assume that each course has a set of inherent topics and therefore relevant words that summarize them. In our content-based recommendation setting those are input features that describe courses. We are concerned in discovering the parameter vector of users, i.e., weights over topic words that denote user preferences on courses. In order to infer this user vector, we rely on supervised machine learning algorithms thus, we assume grading obtained in college courses as ratings, learn user weights and ratings are predicted for unseen MOOCs courses. Preliminary results show suitability of this approach.

The paper is organized as follows. In Section 2 background is given. In Section 3, our proposal is presented. Section 4 shows experimental results. Finally, Section 5 concludes the paper.

2 Background

2.1 Probabilistic Topic Modeling

Topic models are probabilistic models that have been mainly used to discover topics in a big collection of text documents. They are non supervised learning (Duda et al., 2012) techniques that do not require any prior annotations or labeling of the documents: the topics emerge from the analysis of the original texts (Blei, 2012). To do so, they assume each document is a combination of topics and each topic is a probability distribution over words (Blei et al., 2003). Topic models are a type of graphical model based on Bayesian networks.

The generative process described by a topic model does not make any assumptions about the order of words as they appear in documents. The only information relevant to the model is the number of times words are produced, this is known as the "bag-of-words" assumption (Steyvers and Griffiths, 2007).

There are two main topic models: LDA (Blei et al., 2003) and Probabilistic Latent Semantic Analysis (pLSA) (Hofmann, 1999). In this work we use LDA due to its general model. It is also worth noting that LDA has been previously used in recommendation systems (Romero and Ventura, 2010; Romero et al., 2007; Kuang et al., 2011).

2.2 Topics Modeling using Latent Dirichlet Allocation

Latent Dirichlet allocation (LDA) (Blei et al., 2003) is widely used for identifying topics in a set of documents, building on previous work by Hofmann (Hofmann, 1999). The corresponding graphical model representation is depicted in Figure 1, where each document is represented as a mixture of a fixed number of topics, with topic z receiving weight $\theta_z^{(d)}$ in document d, and each topic is a probability distribution over a finite vocabulary of words, with word i having probability $\phi_i^{(z)}$ in topic z.

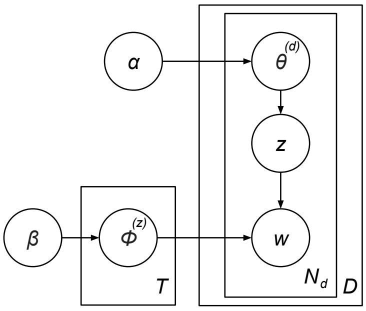


Figure 1: Graphical model for the topic modeling using plate notation

Symmetric Dirichlet priors are placed on $\theta^{(d)}$ and $\phi(j)$, with $\theta^{(d)} \sim$ Dirichlet(α) and $\phi(j) \sim$ Dirichlet(β), where α and β are hyper-parameters that affect the sparsity of these distributions. The hyper-parameter α can be interpreted as a prior observation count for the number of times a topic is sampled in a document, and β as the prior observation count on the number of times words are sampled from a topic before any word from the corpus is observed. This smooths the word distribution in every topic, with the amount of smoothing determined by β. The goal of inference in this model is to identify the values of ϕ and θ, given a corpus of D documents represented by a vocabulary of W words.

In our proposal, each course is a document d that has its related sequence of N_d word tokens, N words in the overall corpus.

2.3 Gibbs Sampling Algorithm

There are many algorithms proposed to obtain the main variables of interest θ and ϕ in the literature, (Hofmann, 1999) used the expectation-maximization (EM) algorithm, this approach suffers from problems involving local maxima of the likelihood function, which has motivated a search for better estimation algorithms like the ones proposed in (Blei et al., 2003; Buntine, 2002; Minka and Lafferty, 2002).

Instead of directly estimating the variables for each document, another approach is the algorithm called "Gibbs sampling" (Griffiths and Steyvers, 2004), which provides a relatively efficient method of extracting a set of topics from a large corpus. Gibbs sampling considers each word token in the text collection in turn, and estimates the probability of assigning the current word token to each topic, conditioned on the topic assignments to all other word tokens. From this conditional distribution, given a document, a topic is sampled and stored as the new topic assignment for this word token. We write this conditional distribution as:

$$P(z_j|z_{N\backslash j}, w_N) = \frac{n^{(w_j)}_{z_j,N\backslash j} + \beta}{n^{(\cdot)}_{z_j,N\backslash j} + W\beta} \cdot \frac{n^{(d_j)}_{z_j,N\backslash j} + \alpha}{n^{(d_j)}_{\cdot,N\backslash j} + T\alpha}$$

where:

$w_N = (w_1, \ldots, w_N)$ are the words in the entire corpus

$z_N = (z_1, \ldots, z_N)$ are the topic assignments of the words

$z_{N\backslash j}$ indicates $(z_1, \ldots, z_{j-1}, z_{j+1}, \ldots, z_N)$

W is the size of the vocabulary

$n^{(w_j)}_{z_j,N\backslash j}$ is the number of times a word wj is assigned to topic z_j

$n^{(\cdot)}_{z_j,N\backslash j}$ is the total number of words assigned to topic z_j

$n^{(d_j)}_{z_j,N\backslash j}$ is the number of times a word in document d_j is assigned to topic z_j

$n^{(d_j)}_{\cdot,N\backslash j}$ is the total number of words in document d_j

From this probability distribution it is possible to make inference, in order to compute conditional probability of topic structure given the observed document. The probability distribution of topics in a document represents a feature vector for that document.

2.4 Recommender Systems

According to (Ricci et al., 2011), recommender systems are software tools and techniques providing items suggestions for a given user. Suggestions provided are aimed at supporting their users in various decision-making processes, such as what items to buy, what music to listen, or what news to read.

As a rule, in a recommendation-system application there are two classes of entities, which we shall refer to as users and items. Users have preferences for certain items and these preferences must be teased out of the data (Rajaraman and Ullman, 2012). The data itself is represented as a utility matrix, giving for each user-item pair, a value that represents what is known about the degree of preference of that user for that item. Values come from an ordered set, e.g., integer $1-5$ representing the number of stars that the users gave as a rating for that item. We assume that the matrix is sparse, meaning that most entries are unknown. An unknown rating implies that we have no explicit information about the user's preference for the item. The goal of a recommendation system is to predict the blanks in the utility matrix.

There are two basic architectures for a recommendation system (Rajaraman and Ullman, 2012):

- Content-based systems focus on properties of items. Similarity of items is determined by measuring the similarity in their properties

- Collaborative-Filtering system focus on the relationship between users and items. Similarity of items is determined by the similarity of the ratings of those items by the users who have rated both items.

In a content-based system, we must construct a profile for each item, which is a record of collections of records representing important characteristics of that item. In simple cases, the profile consist of some characteristics of the item that are easily discovered. For example, in a movie there are the set of actors, the director, the genre of general type of movie. In documents it is not immediately apparent what the values of features should be. There are many kinds of documents for which a recommendation system can be useful. For example, there are many news articles published each day, and we cannot read all of them. A recommendation system can suggest articles on topics a user is interested in. Unfortunately, documents do not tend to have available information giving features. A substitute that has been useful in practice is the identification of words that characterize the topic of a document. An approach is to compute the TF(Term frequency) - IDF(Inverse document frequency) score for words in the document. The ones with the highest scores are the words that characterize the document. In this sense, documents are represented by sets of words. In this paper we have used a different approach which relies on finding document topic information by using topic modeling algorithms such as LDA.

3 Proposal

In order to recommend online courses, each course is considered a document which has a given content. To characterize each course, LDA is used to uncover the semantic structure hidden in the document. Since LDA allow us to get a topic distribution for each course, this output is used as a feature vector for courses (items according to a content-based recommendation setting). A recommendation system is built using item profiles and utility matrices and we treat the problem as one of machine learning. Regard the given data as a training set, and for each user, build a classifier that predicts the rating of all items. The rest of this section describes our main design choices.

courses	*c. features*	*profile user(1)—rating ...*
Calculus	$x_1, \dots x_n$	$\theta_1^{(1)}, \dots, \theta_n^{(1)}$ — **12**
$\cdots$	$\cdots$	$\cdots$
ML(Mooc)	$x_1^{'}, \dots x_n^{'}$	$\theta_1^{(1)}, \dots, \theta_n^{(1)}$—**?**

Table 1: Utility matrix for courses

Consider the utility matrix in Table 1 used to represent a content-based recommendation system. First column contains courses names (college and MOOC's courses). Second column contains feature descriptors for courses. Each row denotes a different course, therefore each course has a different feature vector. Third column shows the user vector profile $\Theta^{(1)}$ for user 1. This vector could comprise user 1 preferences about art, math, biology and social sciences in general. In this same column is also showed user 1 ratings for each course (they are in fact grades obtained in college for user 1, see for instance rating 12 for calculus). Further columns for user 2, user 3 and so on should be added accordingly. Our goal is to predict missing ratings for MOOC's courses (? symbol in last row) for user 1 (user 2, 3, etc.). In order to do so, we should perform the following steps:

- Extract item vectors for courses: item vectors are defined by courses content, i.e., text that describes courses, such as "about the course" information. In order to construct item vectors (features from documents), we rely on Latent Dirichlet Allocation algorithm which extracts topic information from text as probability distribution of words. Since we use a machine learning setting, item vectors are features of a regression/classification problem, which we denote $X = \{X_1, X_2, \dots, X_n\}$.

- Learn user's vector: interests about topic courses can be modeled by user's vector which should be learned for each user. To do so, we use a machine learning approach, all available ratings (grading information in college) are used to train a multilinear regression model (Bishop and others, 2006). The user's vector is therefore the resulting set of parameters (or weights), $\Theta^{(1)} = \{\theta_1^{(1)}, \dots, \theta_n^{(1)}\}$

learned from training data (for instance, all courses and gradings of user 1). There are m (number of users) set of parameters. In a multi-linear regression algorithm we want to find the values for Θ, that minimize the cost function:
$J(\Theta_0, \Theta_1, \dots, \Theta_n) = \frac{1}{2m}\sum_{i=1}^{m}(h_\Theta(x^{(i)}) - y^i)^2$

We define an hypothesis:
$h_\Theta(x) = \Theta^T x = \Theta_0 x_0 + \Theta_1 x_1 + \Theta_2 x_2 + \ldots + \Theta_n x_n$
Where $\Theta_0, \Theta_1, \dots, \Theta_n$ are the parameters we want to predict minimizing the cost function. One way to minimize the cost function is by using gradient descent method, where each iteration of gradient descent makes the parameters θ_j come closer to the optimal values that will minimize the cost function $J(\theta)$.

```
For $n > 1$
Repeat {
$\Theta_j := \Theta_j - \alpha\frac{1}{m}\sum_{i=1}^{m}(h_\Theta(x^{(i)}) - y^i){x_j}^{(i)}$
(simultaneously update $\Theta_j$ for j = 0, ...n) }
```

- Given item and user vectors the goal is to predict a rating R_C for a MOOC course C with feature vector X_C for user U, i.e., user vector profile $\Theta^{(U)}$, the resulting predicted rating is given by:

$$R_C = X_C^T \Theta^{(U)}$$

An overview of the recommendation system is depicted in Figure 2 where we estimate the ratings for a student and to recommend a course we consider a "top-10 best recommendations" approach thus, each student get always 10 recommended courses. Those are the most related MOOCs to courses in which a student get the 10 lowest grades.

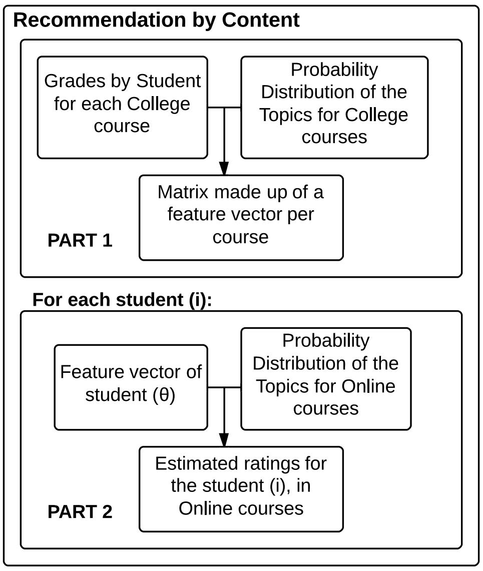

Recommendation by Content
Grades by Student for each College course
Probability Distribution of the Topics for College courses
Matrix made up of a feature vector per course
PART 1
For each student (i):
Feature vector of student (θ)
Probability Distribution of the Topics for Online courses
Estimated ratings for the student (i), in Online courses
PART 2

Figure 2: Block diagram for the recommendation system

4 Experimental Results

This section shows preliminary experimental results conducted on real world data sets. Courses and users grading information where extracted from a Peruvian university. Some MOOC's courses were extracted from Coursera, the following categories were considered: "business and management", "computer science - artificial intelligence", "computer science - software engineering", "computer science - systems and security", "computer science - theory", "mathematics", "statistics and data analysis". The most significant information from each course is given by "Introduction", "About the Course" and "FAQ" sections.

All extracted information has been preprocessed according to the following process: remove non ASCII characters, strip HTML tags, remove special strings, remove multiple spaces and blank lines.

After that we built a corpus further used by the LDA algorithm. The number of Coursera courses considered was 69, while the number of college courses was 43, which gives rises to 112 courses. The topic modeling algorithm used the gibbs sampling inference procedure and according to (Blei, 2012) we set parameters $\alpha = 50/T$, $\beta = 0.01$. The number of iterations was chosen to be large enough to guarantee convergence, $N = 200$.

To measure performance, accuracy was considered by counting the number of correct matches between college courses and Coursera courses. Figure 3 illustrates the impact of the number of topics T in the topic model. A higher accuracy is achieved when we use a higher number of topics, then we set the number of topics $T =$ number of Coursera courses because of the precision.

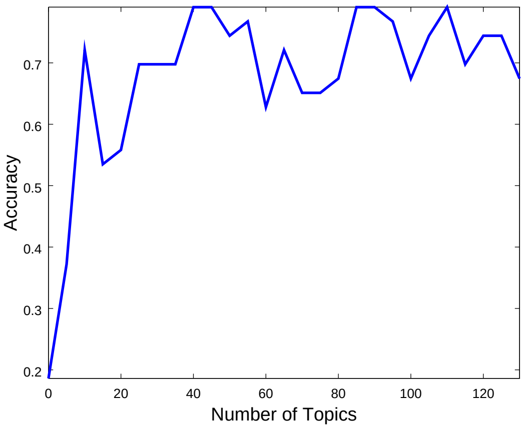


Figure 3: Accuracy of the recommendation system according to the number of topics, a better precision and also efficiency is obtained when the number of topics is equal to the number of coursera courses, T =69

The goal of our proposal is to recommend courses for students who have received low grades in college therefore, we are using grades as ratings. To keep a recommendation system setting, we have decided to invert grading information thus, 20 grade turns out 0 rating and viceversa (this step might not be necessary in other recommendation systems). Mean normalization is also used to get a more reliable recommendation for students with few grades available, for instance, first year students.

For testing, we define a variable "top-N" which denotes the number of courses to recommend. For instance, for student "a" we recommend the "top-N" courses from Coursera where he/she has gotten the greatest ratings. In Figure 4, the x-axis denotes several values for "top-N", and the y-axis denotes accuracy obtained. An cccuracy over 0.6 is achieved for "top-N" greater than or equal to 10.

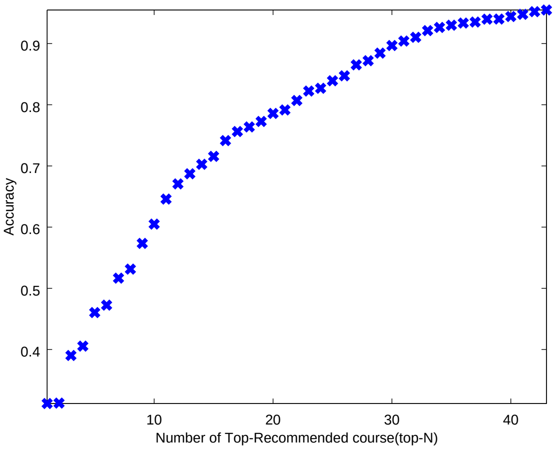


Figure 4: Recommendation system accuracy according to the number of recommended courses

In Figure 5, a comparison between ratings of "coursera courses" and "college courses" for one student is showed. We intend to show proximity of predicted data (ratings on "coursera courses") and provided data (ratings on college courses).

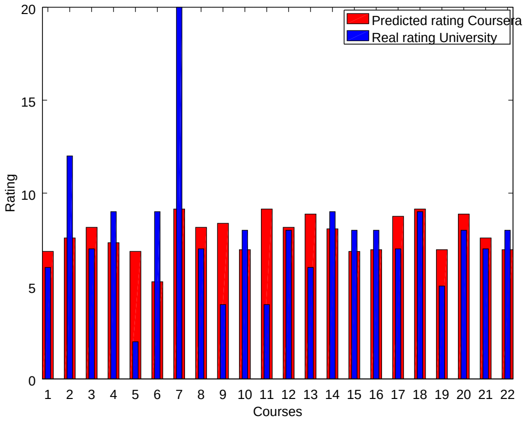


Figure 5: Comparison chart of ratings for one student

5 Conclusion

We have introduced a novel approach for recommending online courses that combines the probabilistic topic model LDA and content-based recommendation systems. In short, we use a machine learning approach where LDA allow us to extract feature descriptors from courses, rating prediction in this setting is performed by inferring user profile parameters using multilinear regression. Preliminary experimental results show that our algorithm performs well when compared to a similar approach based on cosine similarity with LDA.

Although we have focused on MOOCs as source of recommendation content, nothing prevent us from using this approach beyond such domain. In fact, further domains can be included by performing feature topic extraction. Future work will be addressed to investigate scalability issues. In this sense, topic models such as LDA, have scalable versions available. For instance, a MapReduce implementation is given in the Apache Mahout library[1]. There are also scalable versions for multilinear regression.

References

Christopher M Bishop et al. 2006. *Pattern recognition and machine learning*, volume 1. springer New York.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March.

David M Blei. 2012. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

Wray Buntine. 2002. Variational extensions to em and multinomial pca. In *Machine Learning: ECML 2002*, pages 23–34. Springer.

Richard O Duda, Peter E Hart, and David G Stork. 2012. *Pattern classification*. John Wiley & Sons.

Gerhard Fischer. 2014. Beyond hype and underestimation: identifying research challenges for the future of moocs. *Distance Education*, 35(2):149–158.

Thomas L Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235.

Choochart Haruechaiyasak and Chaianun Damrongrat. 2008. Article recommendation based on a topic model for wikipedia selection for schools. 5362:339–342.

Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM.

Wei Kuang, Nianlong Luo, and Zilei Sun. 2011. Resource recommendation based on topic model for educational system. 2:370–374, Aug.

Fred G. Martin. 2012. Will massive open online courses change how we teach? *Commun. ACM*, 55(8):26–28, August.

Thomas Minka and John Lafferty. 2002. Expectation-propagation for the generative aspect model. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pages 352–359. Morgan Kaufmann Publishers Inc.

Anand Rajaraman and Jeffrey David Ullman. 2012. *Mining of massive datasets*. Cambridge University Press, Cambridge.

Francesco Ricci, Lior Rokach, and Bracha Shapira. 2011. *Introduction to recommender systems handbook*. Springer.

C. Romero and S. Ventura. 2010. Educational data mining: A review of the state of the art. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 40(6):601–618, Nov.

Cristbal Romero, Sebastin Ventura, JoseAntonio Delgado, and Paul De Bra. 2007. Personalized links recommendation based on data mining in adaptive educational hypermedia systems. 4753:292–306.

Er Sadikov and Ivan Bratko. 2011. Recommending videolectures with linear regression.

Mark Steyvers and Tom Griffiths. 2007. Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7):424–440.

[1] https://mahout.apache.org/

ANIMITEX project:
Image Analysis based on Textual Information

Hugo Alatrista-Salas and **Eric Kergosien** and **Mathieu Roche** and **Maguelonne Teisseire**
TETIS (Irstea, Cirad, AgroParisTech), Montpellier, France
LIRMM (CNRS, Univ. Montpellier 2), Montpellier, France
firstname.lastname@teledetection.fr

Abstract

With the amount of textual data available on the web, new methodologies of *knowledge extraction* domain are provided. Some original methods allow the users to combine different types of data in order to extract relevant information. In this context, this paper draws the main objectives of the ANIMITEX project which combines spatial and textual data. The data preprocessing step is detailed.

Keywords: Knowledge extraction, Text mining, Satellite images, Spatial feature identification

1 Aims of the ANIMITEX project

A lot of high resolution satellite data are now available. This raises the issue of fast and effective satellite image analysis as it still requires a costly human implication. In this context, remote sensing approaches enable to tackle this challenge. The exploratory and ambitious ANIMITEX project[1] aims at processing massive and heterogeneous textual data (i.e. *big data* context) in order to provide crucial information to enrich the analysis of satellite images.

The large amount of data are associated to a temporal repetitivity that increases. For instance today around ten images are available per year (e.g. SPOT, Landsat), and in 3 years, one image every 5 days (based on Sentinel-2 satellites) will be available.

The ANIMITEX project has many application areas such as image annotation (Forestier et al. 2010). For instance, identifying the precise type of culture or the function of a building is not always possible with the only use of images. Nevertheless, textual data could contain this kind of information and give additional meaning to the images. The development of approaches based on image/text matching becomes crucial in order to complete image analysis tasks (Alatrista-Salas and Béchet 2014). It also enables a better classification of data.

Moreover, image-text matching will enrich Information Retrieval (IR) methods and it will provide users a more global context of data (Sallaberry et al. 2008). This can be crucial for the decision maker in the context of land-use planning projects that have to take into account opinions of experts related to a territory (managers, scientists, associations, specialized companies, and so forth).

In the context of the ANIMITEX project, we plan to investigate two specific scenarios: (i) The construction of a road on the north of Villeveyrac (city close to Montpellier, France), (ii) A port activity area, called *Hinterland*, in Thau area (near to Sète, France). The aim of this case studies is to enrich images with information present in documents, e.g. the opinions extracted in newspapers about land-use planning.

The section 2 describes the proposed data preprocessing step. The section 3 details the partners involved in the project.

[1] http://www.lirmm.fr/∼mroche/ANIMITEX/ (web site in French)

2 Data preprocessing process

The current work focuses on adapting of Natural Language Processing (NLP) techniques for recognition of Spatial Features (SF) and thematic/temporal information (Gaio et al. 2012; Maurel et al. 2011). In the proposed approach, SF appearing in a text, are composed of at least one Named Entity (NE) and one or more spatial indicators specifying its location (Lesbegueries et al. 2006). For this, a set of articles (i.e. 12000 documents) concerning Thau region between the years 2010 and 2013 has been acquired. A second part of the data set is composed of raster files (image mosaics Pleiades - spatial resolution 2x2 m - 4 spectral bands) covering all regions of the Thau lagoon (See Figure 1). Satellite images are available via the GEOSUD Equipex[2].

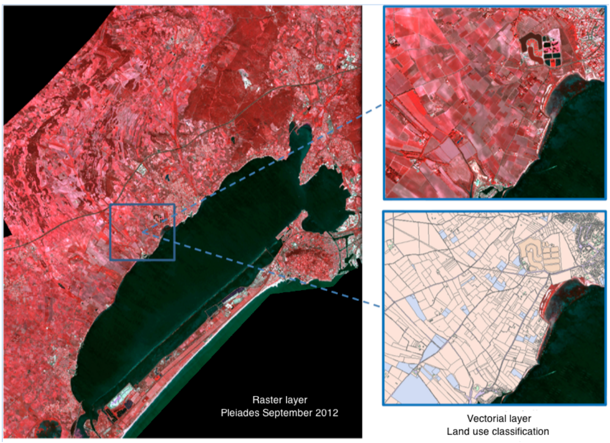


Figure 1: Mosaic images Pleiades around the Thau basin, images on the right represent the superposition of a vector classification on the raster file.

A detailed classification of the land occupation is currently in progress. It will lead to a digital vector layer where each SF (represented by a polygon) belongs to a class of specific land use. The nomenclature of this classification is organized into four hierarchical levels (See Figure 2). Moreover we investigate multi-scale information associated with different levels of classification of satellite images.

From this corpus, NLP methods have been applied in order to identify linguistic features concerning spatial, thematic, and temporal information in the documents. The combined use of lexicons and dedicated rules (Gaio et al. 2012) allows us to identify the absolute (e.g., Montpellier) and relative (e.g., south of Montpellier) Spatial Features (ASF and RSF) (Lesbegueries et al. 2006; Kergosien et al. 2014). A first framework based on sequential pattern mining (Cellier et al. 2010) has been proposed to discover relationships between SF (Alatrista-Salas and Béchet 2014). To this end, a two-step process has been defined (See Figure 3).

SF validation: for each identified ASF, we check on external resources if there is a corresponding spatial representation. In particular, we have used layers provided by the IGN[3] (municipalities, roads, railways, buildings, etc.). In addition, if an ASF does not present on IGN ressources, we use gazetteers (Geonames and Open Street Maps) to complet the information. Concerning the representation of RSF, we use spatial indicators of topological order associates to ASF.

Following the scopes proposed in (Sallaberry et al. 2008), the spatial indicators of topological order have been grouped in five categories:

- Proximity: different indicators can be used in relationship of proximity, such as: *near, around, beside, close to, periphery, etc.*.
- Distance: the indicators used in this relationship are of the form: x *km,* x *miles, etc.*. Two representations are then proposed in our approach: 1) calcul of distance from the centroid of the ASF and construction of a circular buffer of size x from the centroid; 2) regarding the shape of the ASF and building a buffer of size x from the edge of the processed ASF .
- Inclusion: this binary operation allow us to check if an ASF is inside another taking into account indicators such as: center, in the heart, in, inside, etc.
- Orientation: This unary relationship has been broadly studied in the literature. Different approaches have been proposed to identify a cardinal points of an ASF. We have chosen to use the conical model proposed in (Frank 1991). For this, we use the centroid of ASF and we

[2] http://www.equipex-geosud.fr/

[3] Institut National de l'information Gographique et forestire, i.e. National Institute of Geography

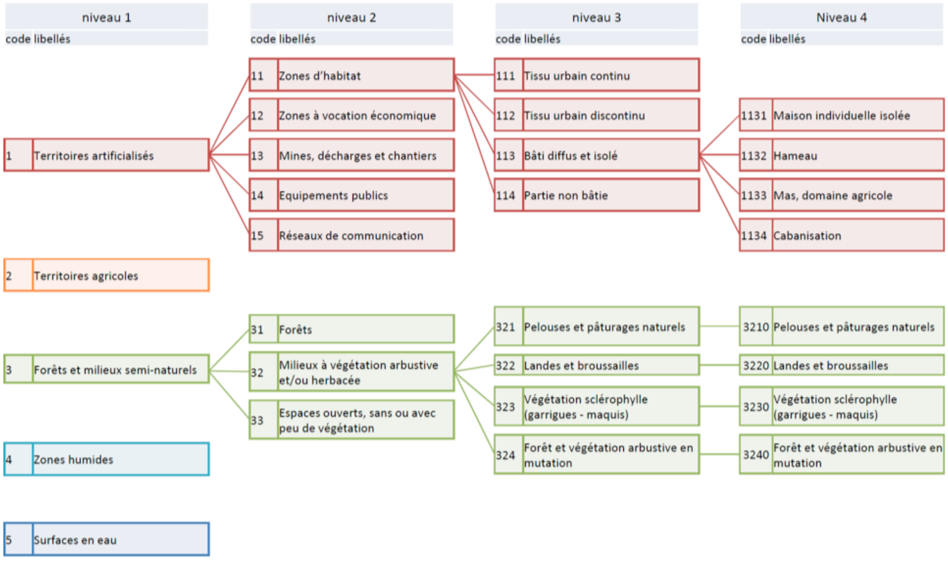

niveau 1
code libellés
niveau 2
code libellés
niveau 3
code libellés
Niveau 4
code libellés
1 Territoires artificialisés
2 Territoires agricoles
3 Forêts et milieux semi-naturels
4 Zones humides
5 Surfaces en eau
11 Zones d'habitat
12 Zones à vocation économique
13 Mines, décharges et chantiers
14 Equipements publics
15 Réseaux de communication
31 Forêts
32 Milieux à végétation arbustive et/ou herbacée
33 Espaces ouverts, sans ou avec peu de végétation
111 Tissu urbain continu
112 Tissu urbain discontinu
113 Bâti diffus et isolé
114 Partie non bâtie
321 Pelouses et pâturages naturels
322 Landes et broussailles
323 Végétation sclérophylle (garrigues - maquis)
324 Forêt et végétation arbustive en mutation
1131 Maison individuelle isolée
1132 Hameau
1133 Mas, domaine agricole
1134 Cabanisation
3210 Pelouses et pâturages naturels
3220 Landes et broussailles
3230 Végétation sclérophylle (garrigues - maquis)
3240 Forêt et végétation arbustive en mutation

Figure 2: Nomenclature of Thau region used to image classification (in French)

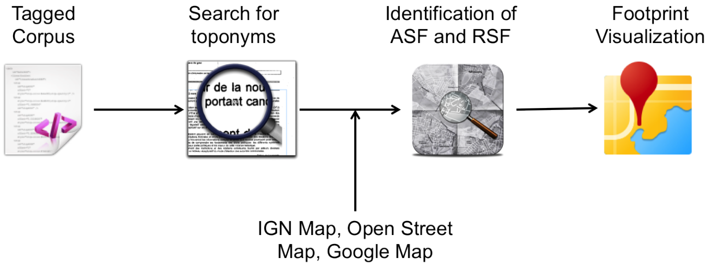

Tagged Corpus
Search for toponyms
Identification of ASF and RSF
Footprint Visualization
IGN Map, Open Street Map, Google Map

Figure 3: From document to footprint definition: the two-step process

build a buffer around. The size of this buffer will be calculated taking into account the surface of the studied ASF. Then we decompose the buffer into four equal areas (forming a ”X”) from the centroid. Each intersection between the buffer and cones thus formed represent the four cardinal points.

- Geometry: geometry relations are built from at least two ASF. These relationships are, for example, the union, the adjacency, the difference or a position of an ASF with respect to other ASF, for example, C between A and B (where A,B and C are ASF), etc.

Representation of the spatial footprint: after the extraction step and spatial representation of the ASF and RSF, the spatial footprint associated with the treated document can be mapped. In this process, two main problems have been identified. The first one is the persistent ambiguity of some NE contained in SF because of some NE (e.g. *Montagnac*) corresponding to several places. To address this issue, a configurable spatial filter based on predefined scenarios has been developed. For example, to identify events related to a specific land-use planning project occurred in a part of the area of the Thau lagoon, only the SF contained in this area will be explored. The second issue is related to the use of external resources and the identification of the spatial representation appropriate to each ASF. Taking into account the spatial indicator (e.g. town, road, etc.) preceding by the toponymic name is a first answer because it allows us to specify the type of the SF and thus take into account the appropriate spatial representation.

Thematic information is identified by semantic resources (i.e. AGROVOC thesaurus, nomenclature resulting of image classifications ...) (Buscaldi et al. 2013).

These linguistic features allow us to identify specific phenomena in documents (e.g., land-use planning, environmental change, natural disasters, etc.). The main idea is to link the phenomena identified in images with subjects found in documents during the same period. Overall, the ANIMITEX project allows the users to integrate different information sources, i.e. both types of expressions (texts vs. images). The main objective is to enrich the information conveyed by a text with images and vice versa.

3 Consortium of the project

The multidisciplinary consortium of the project involves three research domains: Computer Science, Geography and Remote Sensing. More precisely, the expertise in remote sensing and complex mining and heterogeneous spatio-temporal data, is one of the foundations of the project.

TETIS (Territories, Environment, Remote Sensing and Spatial Information, Montpellier) aims to produce and disseminate knowledge, concepts, methods, and tools to characterize and understand the dynamics of rural areas and territories, and control spatial information on these systems. LIRMM (Informatics, Robotics and Microelectronics, Montpellier) focuses on knowledge extraction. ICube (Strasbourg) is specialized in image analysis and complex data mining. ICube collaborates with geographers from LIVE laboratory (Image Laboratory, City, and Environment) and specialists in NLP (LiLPa lab – Linguistics, language, speech). These two locations (Montpellier and Strasbourg) constitute a cluster of local skills related to all major aspects of the project. LIUPPA (Pau) includes researchers specializing in Information Extraction (IE) and Information Retrieval (IR). The main work of this partner is about extraction and managment of geographical information. GREYC (Caen) brings researchers in data mining (e.g. mining sequences in order to discover relationships between spatial entities) and NLP. For this aspect, a collaborations with two other labs is developed (LIPN and IRISA).

Acknowledgments

The authors thank Midi Libre (French newspaper) for its expertise on the corpus and all the partners of ANIMITEX project for their involvement. This work is partially funded by Mastodons CNRS grant and GEOSUD Equipex.

References

Alatrista Salas H., Béchet N. Fouille de textes : une approche séquentielle pour découvrir des relations spatiales. In Cergeo Workshop - EGC, 2014

Buscaldi D., Bessagnet M.N., Royer A., Sallaberry C. Using the Semantics of Texts for Information Retrieval: A Concept and Domain Relation-Based Approach. Proceedings of ADBIS (2) - Advances in Databases and Information Systems, pp. 257-266, 2013.

Cellier P., Charnois T., Plantevit M., Crémilleux B. Recursive Sequence Mining to Discover Named Entity Relations Symposium on Intelligent Data Analysis, LNCS, pp. 30-41, 2010.

Forestier G., Puissant A., Wemmert C., Gançarski, Knowledge-based Region Labeling for Remote Sensing Image Interpretation Computers, Environment and Urban Systems, Vol. 36(5), pp. 470?480, 2012

Frank A. U. Qualitative spatial reasoning with cardinal directions. In *Seventh Austrian Conference on Artificial Intelligence*, volume 287 of *Informatik-Fachberichte*, pages 157–167. Springer, Berlin Heidelberg, 1991.

Gaio M., Sallaberry C., and Nguyen V.T. Typage de noms toponymiques à des fins d'indexation geéographique. *TAL*, 53(2):143–176, 2012.

Kergosien E., Laval B., Roche M., Teisseire M. Are opinions expressed in land-use planning documents? International Journal of Geographical Information Science, Vol. 28(4), pp.739-762, 2014.

Lesbegueries J., Gaio M., and Loustau P. Geographical information access for non-structured data. In *Proceedings of the 2006 ACM Symposium on Applied Computing*, SAC '06, pages 83–89, New York, NY, USA, 2006.

Maurel D., Friburger N., Antoine J.-Y., Eshkol-Taravella I., and Nouvel D. Casen: a transducer cascade to recognize french named entities. *TAL*, 52(1):69–96, 2011.

Sallaberry C., Gaio M., and Lesbegueries J. Fuzzying gis topological functions for gir needs. In Jones C. B. and Purves R., editors, *5th ACM Workshop On Geographic Information Retrieval*, pages 1–8, 2008.

A case study on Morphological Data from Eimeria of Domestic Fowl using a Multiobjective Genetic Algorithm and R&P for Learning and Tuning Fuzzy Rules for Classification

Edward Hinojosa C.
Dept. Informatics Engineering
Pontifical Catholic University of Peru
Lima, Peru
ehinojosa@pucp.pe

Cesar A. Beltran C.
Dept. Informatics Engineering
Pontifical Catholic University of Peru
Lima, Peru
ebeltran@pucp.pe

Abstract

In this paper, we use fuzzy rule-based classification systems for classify cells of the Eimeria of Domestic Fowl based on Morphological Data. Thirteen features were extracted of the images of the cells, these features are genetically processed for learning fuzzy rules and a method reward and punishment for tuning the weights of the fuzzy rules. The experimental results show that our classifier based on interpretability fuzzy rules has a similar classification rate to that of a non-parametric and non-interpretability method.

1 Introduction

The fuzzy systems were proposed by Zadeh at 1965 (Zadeh, 1965) and they are systems based on the theory of the fuzzy sets and logic fuzzy. A of the most important types of fuzzy systems are the Fuzzy Rule Based Classification Systems (FRBCSs) (Herrera, 2005) (Herrera, 2008). Classification problem is studied in the machine learning, data mining, database, and information retrieval communities with applications in a several domains.

The rules are a paradigm for representing knowledge and they have the capacity to build a linguistic model interpretable to the users. The learning (or automatic generation) and tuning of the fuzzy rules in FRBCSs from data sample is a difficult task (Herrera, 2008). This task can be considered as an optimization or search process that can be managed by using Evolutionary Algorithms (EAs). The Genetic Algorithms (GAs) is one of the most know and highly used of EAs. The FRBCSs are defined as Genetic Fuzzy Rule-Based Systems (GFRBSs) when the GAs are used to learn or tuning FRBCSs. The GFRBSs continue to be researched and used in recent years (Nojima and Ishibuchi, 2013), (Chen et al., 2013), (Jalesiyan et al., 2014).

Generally a FRBCSs is composed of two components (Herrera, 2005), the Knowledge Base (KB) and the Inference Mechanism (IM). The KB is composed of two components too, the Data Base (DB) and the Rule Base (RB). This paper is concerned with the genetic learning of the RB.

The most commonly used approaches for rule learning in FRBCSs using GAs are Pittsburgh, Michigan, Iterative Rule Learning (IRL) and Genetic Cooperative-Competitive Learning (GCCL). In the Pittsburgh approach, each chromosome encodes a set of fuzzy rules, after the genetic process the RB is a better chromosome (De Jong et al., 1993). In the Michigan approach, each chromosome encodes a single rule, after the genetic process the RB is the set of chromosomes or rules of the population (Holland and Reitman, 1978). In the IRL approach, each chromosome encodes a single rule too, but after the genetic process, the better rule is selected and inserted to the RB, this process is repeated iteratively until a condition is satisfied (Gonzalez and Perez, 2012). The GCCL approach is a hybrid of the Pittsburgh and Michigan approaches, the rules or chromosomes cooperate among themselves based on Pittsburgh approach and the rules or chromosomes compete among themselves based on Michigan approach (Giordana and Neri, 1995).

This paper is based in the IRL approach using a

Multiobjective Genetic Algorithms (MOGAs). We use MOGAs because in the process of learning fuzzy rules in FRBCSs are considered two objectives: accuracy and interpretability. This objectives are considered contradictory (Casillas and Carse, 2009) and we search a trade-off of them. The accuracy is measured by the classification rate and the interpretability is measured for many features of the FRBCSs, for example, quantity of the rules or quantity of the conditions of each rule. We use specifically the well-known algorithm called Non-dominated Sorting Genetic Algorithm II (NSGA-II) (Deb et al., 2002). After the learning the fuzzy rules, we use a Reward and Punishment(R&P) method for the tuning the factors or weights of the rules (Nozaki et al., 1996) to improve the accuracy of the FRBCS.

We use the proposed method for classify cells of the Eimeria of Domestic Fowl. The Eimeria genus comprises a group of protozoan parasites that infect a wide range of hosts. A total of seven different Eimeria species infect the domestic fowl, causing enteritis with severe economic losses. We use three groups of morphological features: geometric measures, curvature characterization, and internal structure quantification (Beltran, 2007).

This paper is organized as follows: we present in Section 2 the basic concept of classification and FRBCSs employed in this paper. In Section 3 we describe the genetic algorithm multiobjetivo called NSGA-II used in this paper. The proposed method for learning the RB and tuning the factor of each rule is detailed in Section 4. The Section 5 shows the results of the classification on morphological features of the Eimeria genus. The conclusions of this work are presented in Section 6.

2 Fuzzy Rule Based Classification Systems

Classification problem is studied in the machine learning, data mining, database, and information retrieval communities with applications in a several domains, such as medical (Kumar et al., 2013), target marketing (Yongzhi et al., 2013), biology (Silla and Kaestner, 2013), among others.

Any classification problem has a set of examples $E = \{e_1, e_2, ..., e_p\}$ and a set of classes $C = \{C_1, C_2, ..., C_m\}$, the objective is labeled each example $e_q \in E$ with a class $C_j \in C$. Each e_q is defined by a set of features or characteristics $e_q = \{a_{q1}, a_{q2}, ..., a_{qn}\}$.

A FRCS resolves classification problems using rules usually with the follow structures:

R_i: **IF** V_1 **IS** T_{1l_1} **AND** V_2 **IS** T_{2l_2} **AND** ... **AND** T_n **IS** T_{nl_n} **THEN** Class = C_j **WITH** CF_i

where:

R_i : Index of the fuzzy rule i.

$V_1, V_2, ..., V_n$: Linguistic variables or features of each example e_q.

$T_{1l_1}, T_{2l_2}, ..., T_{nl_n}$: Linguistic terms or fuzzy sets used for representing the class C_j.

C_j : The class of the fuzzy rule R_i.

CF_i : The certainty grade (i.e. rule weight) of the rule R_i.

Usually a FRBCS has two main components (Herrera, 2005): The Knowledge Base (KB) and the Inference Mechanism (IM), these are detailed below:

1. The Knowledge Base: The KB is composed of two components:
   (a) The Data Base: The DB contains the membership functions, fuzzy sets or linguistic terms for each linguistic variable of the classification problem.
   (b) The Rule Base: The RB contains the collection of fuzzy rules representing the knowledge.
2. The Inference Mechanism: The IM is the fuzzy logic reasoning process that determines the outputs corresponding to fuzzified inputs (Lakhmi and Martin, 1998). The most common fuzzy inference method for fuzzy classification problems are the classic and general reasing methods (Cordon et al., 2013). This paper uses the classic method.

3 Non-dominated Sorting Genetic Algorithm Multiobjetive II

Is a new version of the NSGA (Srinivas and Deb, 1994), the NSGA-II was proposed by Deb in 2002 (Deb et al., 2002) and it is computationally more efficient, elitist and doesnt need to define additional parameters.

In the NSGA-II, the population Q_t (size N) is generated using the parent population P_t (size N). After this, the two populations are combined for generating the population R_t (size $2N$). The population R_t is sorted according the dominance of the solutions in different Pareto fronts (Pareto, 1896) and the crowding distance. A new population P_{t+1} (size N) is generated with the bests Pareto fronts F_1, F_2, F_3 and so forth, until the P_{t+1} size equals to the value of N. The solutions in the Pareto fronts under this limit are removed. After P_{t+1} is a new P_t and the process is repeated until a conditions is satisfied. The figure 1 shows the process of evolutions of the solutions in the NSGA-II. More details on NSGA-II can be found at (Deb et al., 2002).

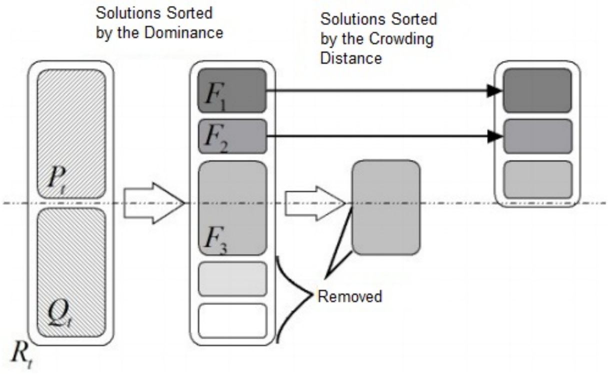

Solutions Sorted by the Dominance
Solutions Sorted by the Crowding Distance
P_t
Q_t
R_t
F_1
F_2
F_3
Removed

Figure 1: Evolutions of the Solutions in the NSGA-II

4 Proposed Method

This section presents the proposed methods for learning fuzzy rules using the IRL approach and a MOGA, and tuning the weights of the fuzzy rules using a R&P method. In the next subsections each method is detailed.

4.1 Learning Fuzzy Rules

The proposed method for learning fuzzy rules is based in the iterative multiobjective genetic method described in (Hinojosa and Camargo, 2012) and uses a MOGA for learning a single fuzzy rule in each iteration of the MOGA. The main difference with the proposed method is the module for defining the order of the class for learning. This method proposed is illustrated in the Figure 2.

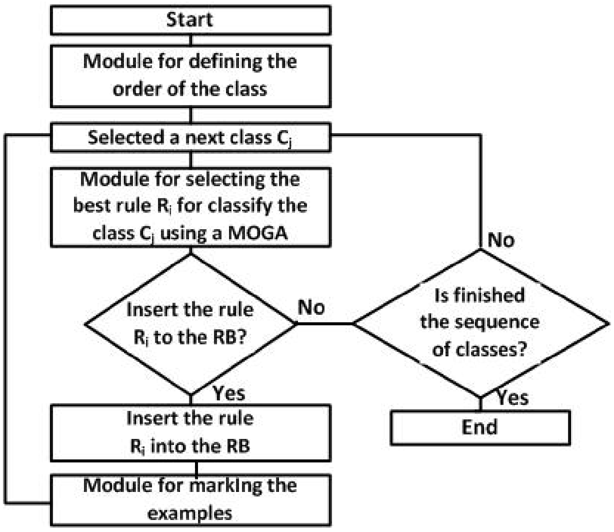

Start
Module for defining the order of the class
Selected a next class C_j
Module for selecting the best rule R_i for classify the class C_j using a MOGA
Insert the rule R_i to the RB?
No
Yes
Insert the rule R_i into the RB
Module for marking the examples
Is finished the sequence of classes?
No
Yes
End

Figure 2: Proposed Method for Learning Fuzzy Rules

A set of examples is used as the set of training. The proposed IRL method starts when is defined the order of the class for learning. After that, a class is selected and the module for generate the best rule that used a MOGA is executed. The MOGA considers two objectives for minimization: accuracy and interpretability. The accuracy is determined by the integrity and consistency of each rule (Gonzalez and Perez, 1999) and the interpretability is defined by the quantity of conditions of each rule. When the best rule in the Pareto front improves the rate of classification of the RB, this rule is inserted into the RB, some examples are marked and the process of learning a fuzzy rule starts again. When the best rule in the Pareto front doesnt improve the rate of classification of the RB, the process verifies that all the sequence of class was learned, if the sequence is not learned a new class is selected and the process of learning a fuzzy rule starts again, else the process finishes and the set of the best rules is the RB.

In the process detailed above all rules has a weight equals to one. These weights can be tuning for improve the rate classification. This tuning is detailed in the next subsection.

4.2 Tunning Weights of the Fuzzy Rules

We use the method proposed in (Nozaki et al., 1996) for this task. This method rewards or increases the weight the a fuzzy rule R_i when a example e_q is correctly classified by this rule according to the next equation:

$$CF_i^{new} = CF_i^{old} + n_1 \left(1 - CF_i^{old}\right) \qquad (1)$$

And this method punishes or decreases the weight of the fuzzy rule R_i when a example e_q is misclassifed by this rule according to the next equation:

$$CF_i^{new} = CF_i^{old} - n_2 CF_i^{old} \qquad (2)$$

In the experimental study detaild in Section 5 we used the values n1=0.001 and n2=0.1 and the tuning procedure for 500 iterations.

5 Experimental Study

The experimental study is aimed to show the application of the proposed method and the comparation with the classification with non-parametric method for classifying cells of the Eimeria of Domestic Fowl based on Morphological Data. The Emeira genus comprises a group of protozoan parasites that infect a wide range of hosts, seven different Emeira species infect the domestic fowl, causing enteritis with several economic losses. This protozoan morphology was represented by 13 features: mean of curvature, standard deviation of curvature, entropy of curvature, major axis (lenght), minor axis (width), symmetry through major axis, symmetry through minor axis, area, entropy of internal structure, second angular moment, contrast, inverse difference moment, entropy of co-occurrence matrix; these features are used as the input pattern for the classification process.

The Table 1 shows the class and the number of examples or instances of each class. More detail how the features were extracted or about the Eimeria genus can be found at (Beltran, 2007).

The Table 2 shows the parameters for learning fuzzy rules and the NSGA-II (the MOGA used in this paper).

After the process of learning fuzzy rules, the process of tuning the weights starts. The Table 3 shows the result of the classification or dispersion matriz after the process tuning the weights.

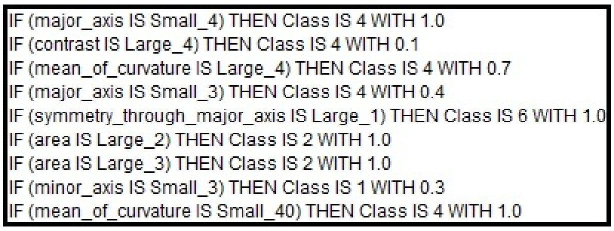
After the proposed method, the result is a set of rules similar of the set shows in the Figure 3. This rules has a high level of interpretability for the expert users.

Class Number	Class Name	# of Examples
1	E. acervulina	636
2	E. maxima	321
3	E. brunetti	418
4	E. mitis	757
5	E. praecox	747
6	E. tenella	608
7	E. necatrix	404

Table 1: Distribution of Classes

Parameter	Value
Size the population	50.0
Crossover rate	1.0
Mutation rate	0.2
Number of generations	500.0
Mark value	0.3

Table 2: Parameters of the Proposed Method



```
IF (major_axis IS Small_4) THEN Class IS 4 WITH 1.0
IF (contrast IS Large_4) THEN Class IS 4 WITH 0.1
IF (mean_of_curvature IS Large_4) THEN Class IS 4 WITH 0.7
IF (major_axis IS Small_3) THEN Class IS 4 WITH 0.4
IF (symmetry_through_major_axis IS Large_1) THEN Class IS 6 WITH 1.0
IF (area IS Large_2) THEN Class IS 2 WITH 1.0
IF (area IS Large_3) THEN Class IS 2 WITH 1.0
IF (minor_axis IS Small_3) THEN Class IS 1 WITH 0.3
IF (mean_of_curvature IS Small_40) THEN Class IS 4 WITH 1.0
```

Figure 3: Proposed Method for Learning Fuzzy Rules

We compared the proposed method with the method non-parametric classifier proposed in (Beltran, 2007) with the same set of examples. The Table 4 shows the classification rate by each class of both classifiers. These results shows that the proposed method (PM) has a similar rate classification (overall 77.17) that the non-parametric method (NPM) (overall 80.24), but with a high degree of interpretability. The non-parametric method does not consider the interpretability.

6 Conclusions

In this article, we proposed a iterative multiobjective genetic method to learn fuzzy classification rules. The fuzzy rules are learned in each iteration depend of the sequencia of class. After that, the weights of the each fuzzy rules are tuned using a R&P method. The results obtained have indicated that FRBCSs

	ACE	**MAX**	**BRU**	**MIT**	**PRA**	**TEN**	**NEC**
ACE	**85.06**	0.00	0.00	2.38	0.13	0.16	7.18
MAX	0.00	**98.44**	0.72	0.00	0.00	0.00	0.00
BRU	0.00	1.56	**87.08**	0.00	6.29	1.48	0.74
MIT	1.73	0.00	0.00	**86.79**	4.95	1.64	4.70
PRA	2.67	0.00	5.50	6.87	**69.34**	10.53	24.01
TEN	7.86	0.00	6.70	1.19	16.06	**82.73**	37.62
NEC	2.67	0.00	0.00	2.77	3.21	3.45	**25.74**

Table 3: Results of the Proposed Method

Class Name	PM	NPM
E. acervulina	85.06	87.70
E. maxima	98.44	96.12
E. brunetti	87.08	94.98
E. mitis	86.79	86.27
E. praecox	69.34	64.46
E. tenella	82.73	76.53
E. necatrix	25.74	55.60

Table 4: The PM vs. NPM

have better interpretability and similar accuracy than a non-parametric method for classify the Eimeria of domestic fowl.

References

Zadeh L. A. 1965. Fuzzy Sets. *Information and Control*, 8(3):338–353.

Herrera F. 2005. Genetic fuzzy systems: Status, critical considerations and future directions. *International Journal of Computational Intelligence Research*, 1(1):59–67.

Herrera F. 2008. Genetic fuzzy systems: taxonomy, current research trends and prospects. *Evolutionary Intelligence*, 1(1):27–46.

Nojima, Y. and Ishibuchi, H. 2013. Multiobjective genetic fuzzy rule selection with fuzzy relational rules. *IEEE International Workshop on Genetic and Evolutionary Fuzzy Systems (GEFS)*, 1:60–67

Chen, S.-M, Chang, Y.-C., Pan, J.-S. 2013. Fuzzy Rules Interpolation for Sparse Fuzzy Rule-Based Systems Based on Interval Type-2 Gaussian Fuzzy Sets and Genetic Algorithms. *IEEE Transactions on Fuzzy Systems*, 21(3):412–425.

Jalesiyan, H., Yaghubi, M., Akbarzadeh, T.M.R. 2014. Rule selection by Guided Elitism genetic algorithm in Fuzzy Min-Max classifier. *Conference on Intelligent Systems*, 1:1–6.

De Jong KA, Spears WM, Gordon DF. 1993. Using genetic algorithms for concept learning. *Mach Learn*, 13:161–188.

Holland, J. and Reitman, J. 1978. *Cognitive Systems Based on Adaptive Algorithms*, ACM SIGART Bulletin

Gonzalez, A. and Perez, R. 1999. SLAVE: A genetic learning system based on an iterative approach. *IEEE Transactions on Fuzzy Systems*, 7:176–191.

Giordana, A. and Neri, F. 1995. Search-intensive concept induction. *Evol Comput*, 3:375–416.

Casillas, J. and Carse, B. 2009. Special issue on Genetic Fuzzy Systems: Recent Developments and Future Directions. *Soft Comput.*, 13(5):417–418.

Deb, K. and Pratap, A. and Agarwal, S. and Meyarivan, T. 2002. A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II. *Trans. Evol. Comp.*, 6(2):182–197.

Nozaki, k., Ishibuchi, H., Tanaka, H. 1996. Adaptive fuzzy rule-based classification systems. *IEEE Trans. Fuzzy Systems*, 4(3):238–250.

Beltran, C. 2007. Anlise e reconhecimento digital de formas biolgicas para o diagnstico automtico de parasitas do gnero Eimeria. *PhD. Teses - USP - Brazil.*

Kumar, S.U., Inbarani, H.H., Kumar, S.S. 2013. Bijective soft set based classification of medical data. *International Conference on Pattern Recognition, Informatics and Mobile Engineering*, 1:517–521.

Yongzhi, Ma., Hong Gao, Yi Ding, Wei Liu. 2013. Logistics market segmentation based on extension classification. *International Conference on Information Management, Innovation Management and Industrial Engineering*, 2:216–219.

Silla, C.N. and Kaestner, C.A.A. 2013. Hierarchical Classification of Bird Species Using Their Audio Recorded Songs. *IEEE International Conference on Systems, Man, and Cybernetics*, 1:1895–1900.

Lakhmi C. and Martin N. 1998. *Fusion of Neural Networks, Fuzzy Systems and Genetic Algorithms: Industrial Applications (International Series on Computational Intelligence).* CRC Press

Cordon, O.,Herrera, F., Hoffmann, F., Magdalena, L. 2001. *Genetic Fuzzy Systems: Evolutionary Tuning and Learning of Fuzzy Knowledge Bases* World Scientific

Srinivas, N. and Deb, K. 1994. Multiobjective Optimization Using Nondominated Sorting in Genetic Algorithms. *Evolutionary Computation*, 2:221–248.

Pareto, V. 1896. *Cours d'Economie Politique*. Droz, Genve.

Hinojosa, E. and Camargo, H.A. 2012. Multiobjective genetic generation of fuzzy classifiers using the iterative rule learning. *International Conference on Fuzzy Systems*, 1:1–8.

Gonzalez, A. and Perez, R. 2009. Improving the genetic algorithm of SLAVE. *Mathware and Soft Computing*, 16:59–70.

SIFR Project: The Semantic Indexing of French Biomedical Data Resources

Juan Antonio Lossio-Ventura,
Clement Jonquet
LIRMM, CNRS, Univ. Montpellier 2
Montpellier, France
fName.lName@lirmm.fr

Mathieu Roche,
Maguelonne Teisseire
TETIS, Cirad, Irstea, AgroParisTech
Montpellier, France
fName.lName@teledection.fr

Abstract

The Semantic Indexing of French Biomedical Data Resources project proposes to investigate the scientific and technical challenges in building ontology-based services to leverage biomedical ontologies and terminologies in indexing, mining and retrieval of French biomedical data.

1 Introducción

Hoy en día la gran cantidad de datos disponibles en línea suele componerse de texto no estructurado, por ejemplo reportes clínicos, informes de reportes adversos, historiales clínicos electrónicos (Lossio-Ventura et al., 2013). Regularmente estos textos son escritos usando un lenguaje específico (expresiones y términos) usados por una comunidad. Es por eso existe la necesidad de formalizar e indexar términos o conceptos técnicos. Lo cual implica un gran consumo de tiempo.

Los términos relevantes son útiles para obtener una mayor comprensión de la estructura conceptual de un dominio. Estos pueden ser: (i) términos de una sola palabra (sencillo a extraer), o (ii) términos de varias palabras (difícil). En el ámbito biomédico, hay una gran diferencia entre los recursos existentes (ontologías) en inglés y francés. En Inglés hay cerca de 7 000 000 de términos asociados a 6 000 000 de conceptos, tales como los de UMLS[1] o BioPortal (Noy et al., 2009). Mientras que, en francés sólo hay alrededor de 330 000 términos asociados a 160 000 conceptos (Neveol et al., 2014). Por lo tanto, hay una necesidad de enriquecer terminologías u ontologías en francés. Por lo tanto, nuestro trabajo se compone de dos pasos principales: (i) la extracción de términos biomédicos, y (ii) el enriquecimiento de ontologías, con el fin de poblar ontologías con los términos extraídos.

El artículo es organizado como sigue. Primero discutimos sobre la sobre la metodología puesta en marcha para este proyecto en la Sección 2. La evaluación de la precisión es presentada en la Sección 3 seguida de las conclusiones en la Sección 4.

2 Metodología

Nuestro trabajo se divide en dos procesos principales: (i) la extracción de términos biomédicos, y (ii) el enriquecimiento de ontologías, explicados a continuación.

2.1 Extracción Automática de Términos Biomédicos

La extracción de términos es una tarea esencial en la adquisición de conocimiento de un dominio. En este trabajo presentamos las medidas creadas para este objetivo. Medidas que se basan en varios criterios como lingüístico, estadístico, grafos y web para mejorar el resultado de extracción de términos biomédicos. Las medidas presentadas a continuación son puestas a disposición de la comunidad, bajo la aplicación llamada BIOTEX (Lossio-Ventura et al., 2014).

2.1.1 Lingüística

Estas técnicas intentan recuperar términos gracias a la formación de patrones. La idea principal es la construcción de reglas para describir las estructuras de los términos de un dominio mediante el uso de características ortográficas, léxicas o morfo-sintácticas. La idea principal es la construcción de reglas, normalmente de forma manual, que describen las estructuras comunes de términos para ciertos campos. En muchos casos también, diccionarios conteniendo términos técnicos (e.g., prefijos, sufijos y acrónimos específicos) son usados para ayudar a extraer

[1] http://www.nlm.nih.gov/research/umls

términos (Krauthammer et al., 2004).

2.1.2 Estadística

Las técnicas estadísticas se basan en la evidencia presentada en el corpus a través de la información contextual. Tales enfoques abordan principalmente el reconocimiento de términos generales (Van Eck et al., 2010). La mayoría de medidas se basan en la frecuencia. La mayor parte de trabajos combinan la información lingüística y estadística, tal es el caso de *C-value* (Frantzi et al., 2000) combina la información estadística y lingüística tanto para la extracción de términos de varias palabras como de términos largos y anidados. Es la medida más conocida en la literatura. En el trabajo de (Zhang et al., 2008), demostraron que *C-value* obtiene los mejores resultados comparado a otras medidas. Además del inglés, *C-value* también ha sido aplicado a otros idiomas tales como japonés, serbio, esloveno, polaco, chino (Ji et al., 2007), español (Barrón-Cedeno et al., 2009), árabe. Es por eso, en nuestro primer trabajo (Lossio-Ventura et al., 2013), la modificamos y adaptamos para el francés.

A partir de *C-value*, hemos creados otras medidas, como *F-TFIDF-C*, *F-OCapi*, *C-OKapi*, *C-TFIDF* (Lossio-Ventura et al., 2014), estas medidas obtienen mejores resultados que *C-value*. Finalmente una nueva medida basada en la información lingüística y estadística es *LIDF-value* (Lossio-Ventura et al., 2014) (patrones **L**ingüísticos, **IDF**, and C-*value* information), que mejora con gran diferencia los resultados obtenidos por las medidas antes citadas.

2.1.3 Grafos

El modelo de grafos es una alternativa al modelo de información, muestra claramente las relaciones entre los nodos gracias a las aristas. Gracias a los algoritmos de centralidad se puede aprovechar los grupos de información en grafos. Existen aplicaciones de grafos para la Recuperación de Información (RI) en el contexto de las redes sociales, de colaboración y sistemas de recomendación (Noh et al., 2009).

Una medida basada en grafos creada para este proceso es *TeRGraph* (Lossio-Ventura et al., 2014) (Terminology Ranking based on Graph information). Esta medida tiene como objetivo mejorar la precisión de los primeros k términos extraídos después de haber aplicado *LIDF-value*. El grafo es construido con la lista de términos obtenidos con *LIDF-value*, donde los nodos representan los términos relacionados con otros términos gracias a la co-ocurrencia en el corpus.

2.1.4 Web

Diferentes estudios de Web Mining se enfocan en la similitud semántica, relación semántica. Esto significa para cuantificar el grado en el que algunas palabras están relacionadas, teniendo en cuenta no sólo similitud sino también cualquier posible relación semántica entre ellos. La primera medida web creada fue *WebR* (Lossio-Ventura et al., 2014), finalmente la mejora llamada *WAHI* (Lossio-Ventura et al., 2014) (**W**eb **A**ssociation based on **H**its **I**nformation). Nuestra medida basada en la Web tiene por objetivo volver a clasificar la lista obtenida previamente con *TeRGraph*. Demostramos con esta medida que la precisión de los k primeros términos extraídos superan los resultados de las medidas arriba mencionadas (ver Sección 3).

2.2 Enriquecimiento de Ontologías

El objetivo de este proceso es enriquecer las terminologías u ontologías con los términos nuevos extraídos en el proceso anterior. Los tres grandes pasos a seguir en este proceso son:

(1) **Determinar si un término es polisémico:** con la ayuda del Meta-Learning, hemos podido predecir con una confianza de 97% si un término es polisémico. Esta contribución será valorizada en la conferencia ECIR 2015.

(2) **Identificar los posibles significados si el término es polisémico:** es nuestro trabajo actual, con la ayuda de clustering, clustering sobre los grafos tratamos de resolver este problema.

(3) **Posicionar el término en una ontología.**

3 Experimentaciones

3.1 Datos, protocolo y validación

En nuestros experimentos, hemos usado el corpus estándar GENIA[2], el cual es compuesto de 2 000 títulos y resúmenes de artículos de revistas que han sido tomadas de la base de datos Medline, contiene más de 400 000 palabras. GENIA corpus contiene expresiones lingüísticas que se refieren a entidades con interés en biología molecular tales como proteínas, genes y células.

[2] `http://www.nactem.ac.uk/genia/genia-corpus/term-corpus`

3.2 Resultados

Los resultados son evaluados en términos de *precisión* obtenidos sobre los primeros k términos extraídos ($P@k$) para las medidas propuestas y las medidas base (referencia) para la extracción de términos compuestos de varias palabras. En las subsecciones siguientes, limitamos los resultados para la medida basada en grafos con sólo los primeros 8 000 términos extraídos y los resultados para la medida basada en la web con sólo los primeros 1 000 términos.

3.2.1 Resultados lingüísticos y estadísticos

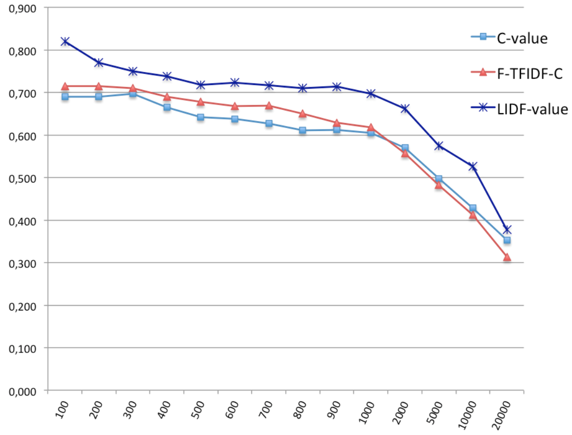


Figure 1: Comparación de la precisión de *LIDF-value* con las mejores medidas de base

3.2.2 Resultados basados en grafos

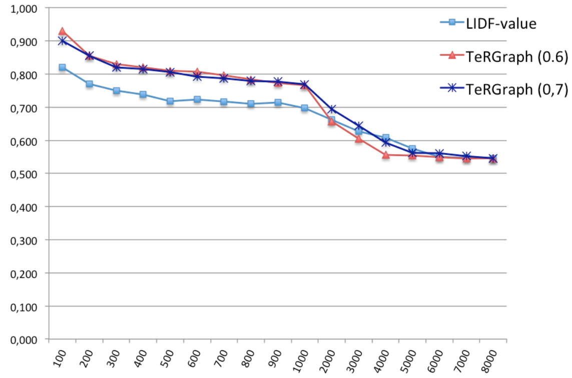


Figure 2: Comparación de la precisión de *TeRGraph* y *LIDF-value*

3.2.3 Resultados basados en la web

4 Trabajo Futuro

Este artículo presenta la metodología propuesta para el proyecto SIFR. Este proyecto consta de dos grandes procesos.

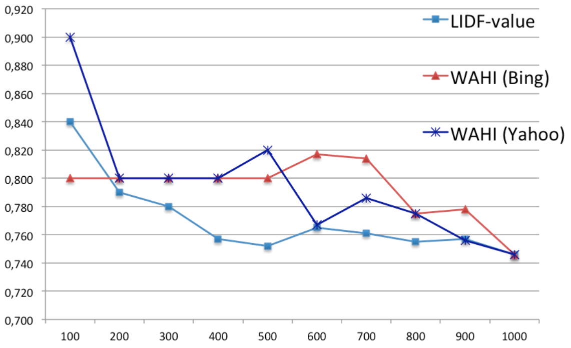


Figure 3: Comparación de la precisión de *WAHI* y *TeRGraph*

El primer proceso *Extracción Automática de Términos Biomédicos*, terminado y siendo valorizado en varias publicaciones citadas anteriormente. En este proceso demostramos que las medidas propuestas mejoran la precisión de la extracción automática de términos en comparación a las medidas más populares de extracción de términos.

El segundo proceso *Enriquecimiento de Ontologías*, a la vez dividio en 3 etapas, es nuestra tarea actual, solo la primera etapa ha sido finalizada. En este proceso buscamos encontrar la mejor posición de un término en una ontología.

Como trabajo futuro, pensamos acabar el segundo proceso. Además, planeamos probar estos enfoques generales sobre otros dominios, tales como ecología y agronomía. Finalmente, planeamos aplicar estos enfoques con corpus en español.

Agradecimientos

Este proyecto es apoyado en parte por la Agencia Nacional de Investigación de Francia bajo el programa JCJC, ANR-12-JS02-01001, así como por la Universidad de Montpellier 2, el CNRS y el programa de becas FINCyT, Perú.

References

Barrón-Cedeno, A., Sierra, G., Drouin, P., Ananiadou, S. 2009. An improved automatic term recognition method for Spanish. *Computational Linguistics, Intelligent Text Processing*, pp. 125-136. Springer.

Frantzi K., Ananiadou S., Mima, H. 2000. Automatic recognition of multiword terms: the C-value/NC-value Method. *International Journal on Digital Libraries*, (3):115-130.

Ji, L., Sum, M., Lu, Q., Li, W., Chen, Y. 2007. Chinese Terminology Extraction Using Window-Based Contextual Information. *Proceedings of the 8th International Conference on Computational Linguistics, Intelligent Text Processing (CICLing07)*, pp. 62-74. Springer-Verlag, Mexico City, Mexico.

Krauthammer, M., Nenadic, G. 2004. Term Identification in the Biomedical Literature. *Journal of Biomedical Informatics*, vol. 37, pp. 512-526. Elsevier Science, San Diego, USA.

Lossio-Ventura, J.A., Jonquet, C., Roche, M., Teisseire M. 2014. BioTex: A system for Biomedical Terminology Extraction, Ranking, and Validation. *Proceedings of the 13th International Semantic Web Conference (ISWC'14)*. Trento, Italy.

Lossio-Ventura, J.A., Jonquet, C., Roche, M., Teisseire M. 2014. Integration of linguistic and Web information to improve biomedical terminology ranking. *Proceedings of the 18th International Database Engineering and Applications Symposium (IDEAS'14)*, ACM. Porto, Portugal.

Lossio-Ventura, J.A., Jonquet, C., Roche, M., Teisseire M. 2014. Yet another ranking function to automatic multi-word term extraction. *Proceedings of the 9th International Conference on Natural Language Processing (PolTAL'14)*, Springer LNAI. Warsaw, Poland.

Lossio-Ventura, J.A., Jonquet, C., Roche, M., Teisseire M. 2014. Biomedical Terminology Extraction: A new combination of Statistical, Web Mining Approaches. *Proceedings of Journées internationales d'Analyse statistique des Données Textuelles (JADT2014)*. Paris, France.

Lossio-Ventura, J.A., Jonquet, C., Roche, M., Teisseire M. 2013. Combining C-value, Keyword Extraction Methods for Biomedical Terms Extraction. *Proceedings of the Fifth International Symposium on Languages in Biology, Medicine (LBM13)*, pp. 45-49, Tokyo, Japan.

Neveol, A., Grosjean, J., Darmoni, S., Zweigenbaum, P. 2014. Language Resources for French in the Biomedical Domain. *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland

Noh, TG., Park, SB., Yoon, HG., Lee, SJ., Park, SY. 2009. An Automatic Translation of Tags for Multimedia Contents Using Folksonomy Networks. *Proceedings of the 32nd International ACM SIGIR Conference on Research, Development in Information Retrieval SIGIR '09*, pp. 492-499. Boston, MA, USA, ACM.

Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D. L., Storey, M., Chute, C.G., Musen, M. A. 2009. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic acids research*, vol. 37(suppl 2), pp 170–173.

Van Eck, N.J., Waltman, L., Noyons, E.CM., Buter, R.K. 2010. Automatic term identification for bibliometric mapping. *Scientometrics*, vol. 82, pp. 581-596.

Zhang, Z., Iria, J., Brewster, C., Ciravegna, F. 2008. A Comparative Evaluation of Term Recognition Algorithms. *Proceedings of the Sixth International Conference on Language Resources, Evaluation (LREC08)*. Marrakech, Morocco.

Mathematical modeling of the performance of a computer system.

Félix Armando Fermín Pérez
Universidad Nacional Mayor de San Marcos
Facultad de Ingeniería de Sistemas e Informática
Lima, Perú
fferminp@unmsm.edu.pe

Abstract

Generally the management of computer systems is manual; autonomic computing by self-management tries to minimize human intervention using autonomic controllers; for this, first the mathematical modeling of the system under study is performed, then is designed a controller that governs the behavior of the system. In this case, the determination of the mathematical model of a web server is based on a black box model by system identification, using data collected and stored in its own log during operation of the computer system under study.

Keywords: autonomic computing, self-management, system identification.

1 Introducción

La tecnología influye en cada aspecto de la vida cotidiana; las aplicaciones informáticas, por ejemplo, son cada vez más complejas, heterogéneas y dinámicas, pero también la infraestructura de información, como la Internet que incorpora grandes cantidades de recursos informáticos y de comunicación, almacenamiento de datos y redes de sensores, con el riesgo de que se tornen frágiles, inmanejables e inseguras. Así, se hace necesario contar con administradores de servicios y de servidores informáticos, experimentados y dedicados, además de herramientas software de monitoreo y supervisión, para asegurar los niveles de calidad de servicio pactados (Fermín, 2012).

Diao et al. (2005) promueven la utilización de la computación autonómica mediante sistemas de control automático en lazo cerrado, reduciendo la intervención humana, que Fox y Patterson (2003) también han identificado como parte principal del problema, debido al uso de procedimientos ad hoc donde la gestión de recursos aún depende fuertemente del control y administración manual. Sobre la computación autonómica, Kephart y Chess (2003) mencionan que la idea es que un sistema informático funcione igual que el sistema nervioso autonómico humano cuando regula nuestra temperatura, respiración, ritmo cardíaco y otros sin que uno se halle siempre consciente de ello, esto es, promueve la menor intervención humana en la administración de la performance de los sistemas informáticos tendiendo hacia la auto-administración de los mismos. Tal auto-administración se caracteriza por las propiedades de auto-configuración (self-configuration), auto-curación (self-healing), auto-optimización (self-optimization) y auto-protección (self-protection).

En la visión de la computación autonómica los administradores humanos simplemente especifican los objetivos de alto nivel del negocio, los que sirven como guía para los procesos autonómicos subyacentes. Así los administradores humanos se concentran más fácilmente en definir las políticas del negocio, a alto nivel, y se liberan de tratar permanentemente con los detalles técnicos de bajo nivel, necesarios para alcanzar los objetivos, ya que estas tareas son ahora realizadas por el sistema autonómico mediante un controlador autonómico en lazo cerrado, que monitorea el sistema permanentemente, utiliza los datos recolectados del propio sistema en funcionamiento, compara estas métricas con las propuestas por el administrador humano, y controla la performance del sistema, por ejemplo, manteniendo el tiempo de respuesta del sistema dentro de niveles prefijados.

De acuerdo con la teoría de control, el diseño de un controlador depende de un buen modelo matemático. En el caso de un sistema informático, primero debe tenerse un modelo matemático para luego diseñar un controlador en lazo cerrado o realimentado, pero sucede que los sistemas informáticos son bastante complicados de modelar ya que su comportamiento es altamente estocástico. Según Hellerstein et al (2004) se ha utilizado la teoría de colas para modelarlos, tratándolos como redes de colas y de servidores, bastante bien pero principalmente en el modelado del comportamiento estacionario y no cuando se trata de modelar el comportamiento muchas veces altamente dinámico de la respuesta temporal de un sistema informático en la zona transitoria, donde la tarea se complica.

De manera que en el presente artículo se trata el modelado matemático de un sistema informático mediante la identificación de sistemas, enfoque empírico donde según Lung (1987) deben identificarse los parámetros de entrada y salida del sistema en estudio, basándose en los datos recolectados del mismo sistema en funcionamiento, para luego construir un modelo paramétrico, como el ARX por ejemplo, con las técnicas estadísticas de autoregresión. La sección 2 describe alguna teoría básica sobre las métricas para el monitoreo de la performance de un sistema informático. En la sección 3 se trata el modelado matemático, y en la sección 4 se describe el experimento realizado; finalmente en la sección 5 se describen las conclusiones y trabajos futuros.

2 Monitoreo de la performance.

En la computación autonómica los datos obtenidos del monitoreo de la performance del sistema en estudio contribuye fundamentalmente en la representación del estado o del comportamiento del sistema, esto es, en el modelo matemático del mismo. Según Lalanda et al. (2013), conocer el estado del sistema desde las perspectivas funcionales y no funcionales es vital para llevar a cabo las operaciones necesarias que permitan lograr los objetivos en el nivel deseado y el monitoreo de la performance permite saber cuan bien lo está logrando. Generalmente los datos de la performance se consiguen vía el log del sistema en estudio, con herramientas de análisis utilizando técnicas de análisis estadísticas, principalmente.

Entre las métricas de performance inicialmente se encontraba la velocidad de procesamiento, pero al agregarse más componentes a la infraestructura informática, surgieron nuevas métricas, siendo las principales las que proporcionan una idea del rendimiento o trabajo realizado en un periodo de tiempo, la utilización de un componente, o el tiempo en realizar una tarea en particular como por ejemplo el tiempo de respuesta. Lalanda et al. (2013) mencionan que las métricas de performance más populares son las siguientes:

- Número de operaciones en punto flotante por segundo (FLOPS), representa una idea del rendimiento del procesamiento, realiza comparaciones entre máquinas que procesan complejos algoritmos matemáticos con punto flotante en aplicaciones científicas.
- Tiempo de respuesta, representa la duración en tiempo que le toma a un sistema llevar a cabo una unidad de procesamiento funcional. Se le considera como una medición de la duración en tiempo de la reacción a una entrada determinada y es utilizada principalmente en sistemas interactivos. La sensibilidad es también una métrica utilizada especialmente en la medición de sistemas en tiempo real, consiste en el tiempo transcurrido entre el inicio y fin de la ejecución de una tarea o hilo.
- Latencia, medida del retardo experimentado en un sistema, generalmente se le utiliza en la descripción de los elementos de comunicación de datos, para tener una idea de la performance de la red. Toma en cuenta no solo el tiempo de procesamiento de la CPU sino también los retardos de las colas durante el transporte de un paquete de datos, por ejemplo.
- Utilización y carga, métricas entrelazadas y utilizadas para comprender la función de administración de recursos y proporcionan una medida de cuan bien se están utilizando los componentes de un sistema y se describe como un porcentaje de utilidad. La carga mide el trabajo realizado por el sistema y usualmente es representado como una carga promedio en un periodo de tiempo.

Existen muchas otras métricas de performance:
- número de transacciones por unidad de costo.

- función de confiabilidad, tiempo en el que un sistema ha estado funcionando sin fallar.
- función de disponibilidad, indica que el sistema está listo para ser utilizado cuando sea necesario.
- tamaño o peso del sistema, indica la portabilidad.
- performance por vatio, representa la tasa de cómputo por vatio consumido.
- calor generado por los componentes ya que en sistemas grandes es costoso un sistema de refrigeración.

Todas ellas entre otras más permiten conocer mejor el estado no funcional de un sistema o proporcionar un medio para detectar un evento que ha ocurrido y ha modificado el comportamiento no funcional. Así, en el presente caso se ha elegido inicialmente como métrica al tiempo de respuesta, ya que el sistema informático en estudio es un servidor web, por esencia de comportamiento interactivo, de manera que lo que se mide es la duración de la reacción a una entrada determinada.

3 Modelado matemático.

En la computación autonómica, basado en la teoría de control realimentado, para diseñar un controlador autonómico primero debe hallarse el modelo matemático del sistema en estudio, tal como se muestra en la Figura N° 1. El modelo matemático de un servidor informático se puede hallar mediante dos enfoques: uno basado en las leyes básicas, y otro en un enfoque empírico denominado identificación de sistemas. Parekh et al. (2002) indican que en trabajos previos se ha tratado de utilizar principios, axiomas, postulados, leyes o teorías básicas para determinar el modelo matemático de sistemas informáticos, pero sin éxito ya que es difícil construir un modelo debido a su naturaleza compleja y estocástica; además es necesario tener un conocimiento detallado del sistema en estudio, más aún, cuando cada cierto tiempo se va actualizando las versiones del software, y finalmente que en este enfoque no se considera la validación del modelo.

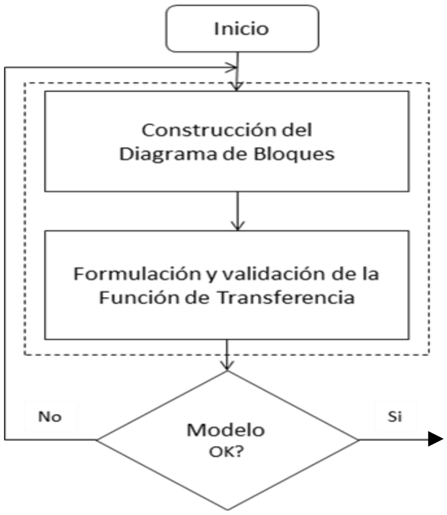

Inicio
Construcción del Diagrama de Bloques
Formulación y validación de la Función de Transferencia
No
Si
Modelo OK?

Figura N° 1. Modelado matemático basado en la teoría de control. (Adaptado de Hellerstein, 2004).

En contraste, según Ljung (1987) la identificación de sistemas es un enfoque empírico donde debe identificarse los parámetros de entrada y salida del sistema en estudio, para luego construir un modelo paramétrico, como el ARX por ejemplo, mediante técnicas estadísticas de autoregresión. Este modelo o ecuación paramétrica relaciona los parámetros de entrada y de salida del sistema de acuerdo a la siguiente ecuación:

$$y(k+1) = Ay(k) + Bu(k) \tag{1}$$

donde	y(k)	: variable de salida
	u(k)	: variable de entrada
	A, B	: parámetros de autoregresión
	k	: muestra k-ésima.

Este enfoque empírico trata al sistema en estudio como una caja negra, de manera que no afecta la complejidad del sistema o la falta de conocimiento experto, incluso cuando se actualicen las versiones del software bastaría con estimar nuevamente los parámetros del modelo. Así, para un servidor web Apache, la ecuación paramétrica relaciona el parámetro entrada, Max Clients (MC), un parámetro de configuración del servidor web Apache que determina el número máximo de conexiones simultáneas de clientes que pueden ser servidos; y el parámetro de salida Tiempo de respuesta (TR), que indica lo rápido que se responde a las solicitudes de servicio de los clientes del servidor, ver Figura N° 2.

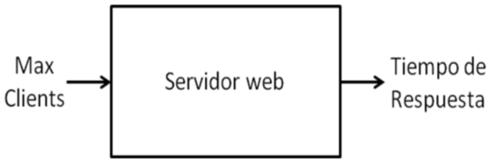


Figura N° 2. Entrada y salida del sistema a modelar. (Elaboración propia).

En (Hellerstein et al, 2004) se propone realizar la identificación de sistemas informáticos, como los servidores web, de la siguiente manera:
1.- Especificar el alcance de lo que se va a modelar en base a las entradas y salidas consideradas.
2.- Diseñar experimentos y recopilar datos que sean suficientes para estimar los parámetros de la ecuación diferencia lineal del orden deseado.
3.- Estimar los parámetros del modelo utilizando las técnicas de mínimos cuadrados.
4.- Evaluar la calidad de ajuste del modelo y si la calidad del modelo debe mejorarse, entonces debe revisarse uno o más de los pasos anteriores.

4 Experimento.

La arquitectura del experimento implementado, se muestra en la Figura Nº 3.

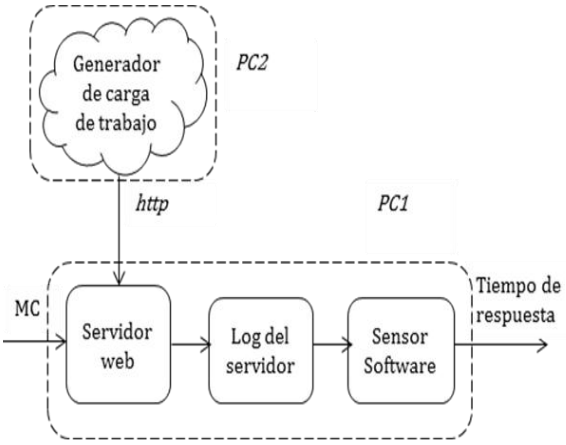


Figura N° 3. Arquitectura para la identificación del modelo de un servidor web (Elaboración propia).

La computadora personal PC1 es el servidor web Apache, asimismo contiene el sensor software que recoge los tiempos de inicio y fin de cada http que ingresa al servidor web, datos que se almacenan en el log del mismo servidor, y luego se realiza el cálculo del tiempo de respuesta de cada http completado en una unidad de tiempo y el tiempo de respuesta promedio del servidor. El servidor web por sí mismo no hace nada, por ello, la computadora personal PC2 contiene un generador de carga de trabajo, que simula la actividad de los usuarios que desean acceder al servidor web; en este caso se utilizó el JMeter una aplicación generadora de carga de trabajo y que forma parte del proyecto Apache.

La operación del servidor web, no solo depende de la actividad de los usuarios, sino también de la señal de entrada MaxClients (MC), que toma forma de una sinusoide discreta variable que excita al servidor web junto con las solicitudes http del generador de carga de trabajo (PC2). De manera que con la actividad del servidor web almacenada en su log, un sensor software calcula los valores de la señal de salida Tiempo de Respuesta (TR), mostrados en la Figura N° 4.

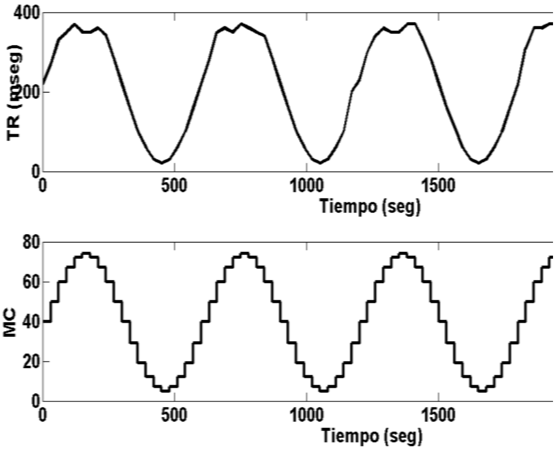


Figura N° 4. Señal de entrada MaxClients MC y señal de salida Tiempo de respuesta TR.

Con los datos de MC y TR obtenidos se estiman los parámetros de regresión A y B de la ecuación paramétrica ARX, haciendo uso de las técnicas de mínimos cuadrados, implementadas en el ToolBox Identificación de Sistemas del Matlab, logrando la siguiente ecuación paramétrica de primer orden:

$$TR(k+1) = 0.06545TR(k) + 4.984MC(k+1)$$

Se puede observar que el Tiempo de respuesta actual depende del Tiempo de respuesta anterior y del parámetro de entrada MaxClients. El modelo hallado es evaluado utilizando la métrica r2, calidad de ajuste, del ToolBox utilizado, que indica el porcentaje de variación respecto a la señal original. En el caso de estudio, el modelo hallado tiene una calidad de ajuste del 87%, lo que se puede considerar como un modelo aceptable. En la Figura N° 5 puede observarse la gran similitud entre la señal de salida medida y la señal de salida estimada.

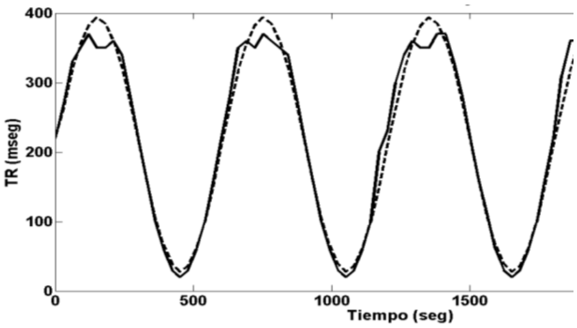


Figura N° 5. TR medida y TR estimada.

5 Conclusiones.

La identificación del comportamiento de un sistema informático tratado como una caja negra es posible, para ello debe simularse el funcionamiento del mismo con hardware y herramientas software como el Jmeter y Matlab, por ejemplo.

El modelo matemático determinado en base a los datos recopilados del log del sistema en estudio se aproxima bastante al modelo real, en este caso se obtuvo un 87% de calidad de ajuste.

El sensor software implementado ha permitido calcular el tiempo de respuesta en base a los datos almacenados en el log del mismo servidor.

De los datos tomados se observa que los sistemas informáticos poseen un comportamiento cercano al lineal solo en tramos por lo que se sugiere experimentar con modelos matemáticos no lineales para comparar la calidad de ajuste de ambos.

Como trabajos futuros se plantea diseñar un controlador autonómico basado en el modelo lineal ARX, aunque más adelante se planteará el diseño de controladores autonómicos no lineales para sistemas informáticos en general, ya que el comportamiento temporal no lineal, hace adecuada la utilización de técnicas de inteligencia artificial como la lógica difusa y las redes neuronales.

6 Referencias bibliográficas.

Diao, Y.; Hellerstein, J.; Parekh, S.; Griffith, R.; Kaiser, G.; Phung, D. (2005) *A Control Theory Foundation for Self-Managing Computing Systems*, IEEE Journal on Selected Areas in Communications, Vol. 23, Issue 12, pp. 2213 – 2222. IEEE. DOI: 10.1109/JSAC.2005.857206.

Fermín F. (2012). *La Teoría de Control y la Gestión Autónoma de Servidores Web*. Memorias del IV Congreso Internacional de Computación y Telecomunicaciones. ISBN 978-612-4050-57-2.

Fox, A.; Patterson, D. (2003). *Self-repairing computers*, Scientific American, Jun2003, Vol. 288 Issue 6, pp. 54-61.

Hellerstein, J.; Diao, Y.; Parekh, S.; Tilbury, D. (2004). *Feedback Control of Computing Systems*. Hoboken, New Jersey: John Wiley & Sons, Inc. ISBN 0-471-26637-X

Kephart, J.; Chess, D. (2003). *The vision of autonomic computing.* Computer, Jan2003, Vol. 36, Issue 1, pp. 41-50. IEEE. DOI: 10.1109/MC.2003.1160055

Lalanda, P.; McCann, J.; Diaconescu, A. (2013). A*utonomic Computing. Principles, Design and Implementation.* London: Springer-Verlag. ISBN 978-1-4471-5006-0

Ljung L. (1987). *System Identification: Theory for the User.* Englewood Cliffs, New Jersey: Prentice Hall.

Parekh, S.; Gandhi, N.; Hellerstein, J.; Tilbury, D.; Jayram, T.; Bigus, J. (2002). *Using control theory to achieve service level objectives in performance management.* Real-Time Systems. Jul2002, Vol. 23, Issue 1/2, pp. 127-141. Norwell, Massachusetts: Kluwer Academic Publishers. DOI: 10.1023/A:1015350520175.

Clasificadores supervisados para el análisis predictivo de muerte y sobrevida materna

Pilar Vanessa Hidalgo León

Universidad Andina del Cusco/San Jerónimo, Cusco-Perú

phidalgo@uandina.edu.pe

Resumen

El presente trabajo se basa en el análisis de los clasificadores supervisados que puedan generar resultados aceptables para la predicción de la muerte y sobrevida materna, según características de pacientes complicadas durante su gestación determinada por los expertos salubristas. Se describe la metodología del desarrollo, las particularidades de la muestra, además los instrumentos utilizados para el procesamiento de los datos. Los resultados de la investigación luego de la evaluación de cada clasificador y entre ellos el que mejores resultados arroja. Los histogramas acerca de cada atributo de las pacientes, y la inclusión en la muestra. Los parámetros determinantes para su correcta clasificación. La comparación de cada resultado entre cada tipo de clasificador dentro de la familia a la que pertenece para después de identificado, implementar, el algoritmo de Naive-Bayes con estimador de Núcleo activado (KERNEL=TRUE), en un software que contribuya a la toma de decisiones certera y respaldada para los profesionales de la salud. En conclusión se encontró un clasificador supervisado que responde positivamente a dar cambio y mejora de la problemática que abarca a la sobrevida materna a pesar de sus complicaciones.

Palabras Clave: Mortalidad materna, clasificadores supervisados, Redes bayesianas, Aprendizaje supervisado.

1 Introducción

El objetivo en el uso de clasificadores supervisados (Araujo, 2006), es construir modelos que optimicen un criterio de rendimiento, utilizando datos o experiencia previa. En ausencia de la experiencia humana, para resolver una disyuntiva que requiere explicación precisa, los sistemas implementados por modelos clasificadores han sido parte importante en la toma de decisiones. A parte, cuando este problema requiere prontitud por su naturaleza, los clasificadores transforman los datos en conocimiento y aportan aplicaciones exitosas.

En el caso de factores de riesgo para la salud materna, existen estudios estadísticos y aplicaciones salubristas para determinarla, mas no integrados simultáneamente como parte de una probabilidad clasificatoria como modelo.

Por ello, este estudio determinará, el clasificador supervisado más eficiente en tiempo y resultado que establezca la diferencia de clases entre pacientes gestantes complicadas durante su embarazo que pueden llegar a presentar síntomas fatales y las que no, así apoyar al personal de salud a tomar la decisión más optima y a prevenir futuras alzas en el índice de mortalidad de su comunidad.

Los problemas que generan el alza de este indicador ya son conocidos y puestos en valor en esta investigación:

"Hoy en día existe suficiente evidencia que demuestra que las principales causas de la muerte materna son la hemorragia posparto, la preclampsia o la sepsis y los problemas relacionados con la presentación del feto. Asimismo, sabemos cuáles son las medidas más eficaces y seguras para tratar estas emergencias obstétricas. Para poder aplicarlas, es necesario que la gestante acceda a un establecimiento de salud con capacidad resolutiva, pero lamentablemente muchas mujeres indígenas no acuden a este servicio por diversas razones, tanto relacionadas con las características geográficas, económicas, sociales y culturales de sus grupos poblacionales, como por las deficiencias del propio sistema de salud.

En los últimos años se han hecho muchos

esfuerzos para revertir esta situación, tanto mediante proyectos promovidos por el estado como ejecutados por organismos no gubernamentales de desarrollo. Estos esfuerzos han tenido, sin embargo, resultados desiguales debido principalmente a la poca adecuación de los proyectos al contexto geográfico y de infraestructura en el que vive gran parte de la población indígena, a sus dificultades económicas para acceder al servicio, su cultura, sus propios conceptos de salud y enfermedad, y su sistema de salud."(Cordero, 2010)

2 Contenido

Problema

¿Cuáles son los clasificadores supervisados que predicen la muerte o la sobrevida materna con mayor efectividad?

Entonces para determinar adecuadamente la efectividad de cada algoritmos nos cuestionamos:

- ¿Cuál es la especificidad, la clasificación correcta y el error absoluto y sensibilidad del clasificador supervisado en relación a los datos de mortalidad materna?
  - Redes neuronales
  - Redes Bayesianas
  - Regresión Logística
  - Arboles de Decisión
  - Algoritmos Basados en Distancias

Mediante la herramienta Weka, se determinó la sensibilidad, la certeza más cercana de cada uno de estos algoritmos, y cuya conclusión sugerirá el más eficiente.

Actualmente la problemática en mortalidad materna es un indicador determinante de desarrollo en los países Latinoamericanos. Siendo no solo un indicador de pobreza y desigualdad sino de vulnerabilidad de los derechos de la mujer. (OMS, 2008)

Limitaciones de la Investigación

Los datos recolectados para este estudio con respecto a pacientes fallecidas que tuvieron complicaciones durante el embarazo fueron 48, ya que el estado del registro de las historias clínicas correspondientes a los casos de fallecimiento no son legibles, ni están conservadas en las mejores condiciones en los archivos de la Dirección Regional de Salud Cusco, esto hace que la muestra no pueda ser nutrida con mayor diversidad de datos. (DIRESA, 2007)

Existe poca investigación acerca del tema relacionado con el uso de clasificadores supervisados y otras correspondientes a diagnósticos médicos que tienen similitud con la muestra pero ninguna que se relacione directamente.

La relevancia de los datos se limito a los antecedentes sobre estudios en mortalidad materna (edad, estado civil, analfabeta, ocupación, procedencia, anticoncepción, entorno (estrato social), controles pre-natales, ubicación domiciliaria, tiempo de demora en atención, atención profesional, antecedentes familiares, espacio intergenésico (en años), paridad (número de hijos), complicaciones no tratadas, fallecimiento), conservando el anonimato de cada paciente.

Objetivos

- Determinar el clasificador supervisado que brinde mejores resultados para el análisis predictivo de muerte y sobrevida materna.

  Luego para lograr este objetivo se debe:

- Determinar la especificidad, la clasificación correcta, el error absoluto, la sensibilidad del clasificador supervisado en relación a los datos de muerte y sobrevida materna.
  - Redes neuronales
  - Redes Bayesianas
  - Regresión Logística
  - Arboles de Decisión
  - Algoritmos Basados en Distancias

Hipótesis General.

Hi: Existen clasificadores supervisados que predicen la muerte o la sobrevida materna con efectividad

H. nula: No existen clasificadores supervisados que predicen la muerte o la sobrevida materna con efectividad

Definición de Variables:

Variable principal:

Clasificadores supervisados

Variables Implicadas:

Variable	Clasificadores supervisados
Dimensión	Las Redes Neuronales, Algoritmos supervisados, Las Redes Bayesianas, Arboles de decisión, Regresión Logística, Algoritmos basados en instancias
Indicador/ Criterios de Medición	Clasificación correcta, Clasificación incorrecta, Sensibilidad,Especificidad, Tiempo de ejecución Mean absolute error Kappa statistic Root mean squared erro, Relative absolute error Root relative squared error
Clase	Numérica discreta
Instrumento	Weka 3.5.7, Explored

Tabla I: Variables implicadas

Metodología de investigación

Cuasi Experimental, Aplicada, Inductiva

Descripción de la muestra y método de recolección

Los datos recolectados en todas 100 historias clínicas (HC) de casos de sobrevida y de casos de muerte materna.

Las características de las gestantes son muy similares entre si y corresponden a la población de la ciudad del Cusco, del archivo en la Red Sur de la Dirección Regional de Salud sobre el control sanitario de mortalidad materna de la Región Cusco.

El número es limitado, pues las historias desde 1992 al 2011, no han sido redactadas ni conservadas en el mejor estado haciendo difícil la tarea de interpretar los datos suficientes para ser analizados.

Estas pacientes no fueron necesariamente atendidas desde el inicio en estos establecimientos, sino que debido a sus complicaciones durante el parto y e embarazo fueron derivadas a las capitales y luego a los establecimientos de mayor capacidad resolutiva para su atención.

Los datos de las historias clínicas que incluyeran:

- Edad
- Estado civil
- Analfabeta
- Ocupación
- Procedencia
- Anticoncepción
- Entorno (estrato social)
- Controles pre-natales
- Ubicación domiciliaria
- Tiempo de demora en atención
- Atención profesional
- Antecedentes familiares
- Espacio intergenésico (en años)
- Paridad (#de hijos)
- Complicaciones no tratadas
- Fallecimiento

Entre 52 sobrevivientes y 48 fallecidas, ambos grupos con similares características, siendo factores determinantes: (Ramírez, 2009)

Ubicación domiciliaria /tiempo demora en atención
Controles > 2: n (49.0/2.0)
PRODECENCIA = rural: s (6.0/1.0)

Técnica e instrumentos de investigación

Se utilizó la herramienta Weka Explorer para la interpretación de los datos.

Las opciones de clasificación supervisada y los algoritmos que propone esta herramienta. (Corso, 2009)

Se evaluaron los siguientes clasificadores supervisados:

- Las Redes Neuronales
  - MultilayerPerceptron
  - RBFNetwork
- Las Redes Bayesianas
  - BayesNet
  - Bayes simple estimator
  - BMA bayes
  - Naive-Bayes
  - BayesNet Kernel
  - Naive-Bayes
  - Discretizacion Supervisada
- Arboles de decisión
  - J48
  - DecisionTable
- Regresión Logística
  - MultiClassClassifier
  - Logistic
- Algoritmos basados en Distancias
  - IBK
  - LWL
  - KStar

Estos resultados fueron comparados con las reglas de clasificación que nos proporcionará los algoritmos de predicción inmediata:

- OneR
- ZeroR

Procedimiento de recolección de datos

Las Historias Clínicas se insertaron en un fichero CSV (delimitado por comas) cuya cabecera contiene las etiquetas de cada atributo, y la última columna se refiere a la clase a la pertenecen.

Con respecto a los indicadores particulares en cada uno de los atributos de los sujetos de la clase tenemos los siguientes valores: Tabla II

- La calidad de la estructuración
- Comparar todas mediciones en cada clasificador por algoritmo.
- Evaluar los resultados.
- Interpretar los resultados.

2.9. Plan de análisis de la información

- Determinación de objetivos
- Preparación de datos
- Selección de datos:
  - o Identificación de las fuentes de información externas e internas y selección del subconjunto de datos necesario.
- Pre procesamiento: estudio de la calidad de los datos y determinación de las operaciones de minería que se pueden realizar.
- Transformación de datos: conversión de datos en un modelo analítico.
- Análisis de datos interpretación de los resultados obtenidos en la etapa anterior, generalmente con la ayuda de una técnica de visualización.
- Asimilación de conocimiento descubierto(Calderón, 2006)
- Minería de datos: tratamiento automatizado de los datos seleccionados con una combinación apropiada de algoritmos. (Ramirez, 2011).

	NEGATIVO	**POSITIVO**	**VALORES**
Edad	Menor a 19 y mayor 35	Entre 19-35	14-48
Estado civil	Soltera	Pareja	Soltera-pareja
Analfabeta	Analfabeta	Primaria	Analfabeta-primaria-secundaria-superior
Ocupación	No remunerada	Remunerada	Remunerada-no remunerada
Procedencia	Rural	Urbana	Rural-urbana
Anticoncepción	No	Si	Si-no
Entorno (estrato social)	Bajo	Medio-alto	Baja-alta
Controles	De 1 a 5	6 a mas	0-12
Ubicación domiciliaria/ tiempo demora en atención	Más de 2 horas	Menos 3 del ee.ss	Menos de 1 hora, 1-2,3-5,6 a mas
Personal de atención profesional	No	Si	No-si
Antecedentes familiares	No	Si	No-si
Espacio intergenésico (en años)	Menos de 2 mayor a 4	Entre 2-4	Primera gesta,1-3,4-6, menos 1
Paridad (#de hijos)	Primípara o mas de 4	Entre 2-4	0-10
Complicaciones no tratadas	Complicaciones antes y durante	Sin complicaciones	No-si
Fallecida	Si	No	No-si

Tabla II: Rango de valores de los atributos en las pacientes de la muestra

Histogramas:

Cada Atributo es evaluado visualmente por los histogramas que arroja Weka (en total 15), por ejemplo con respecto a la EDAD de las pacientes de la muestra.

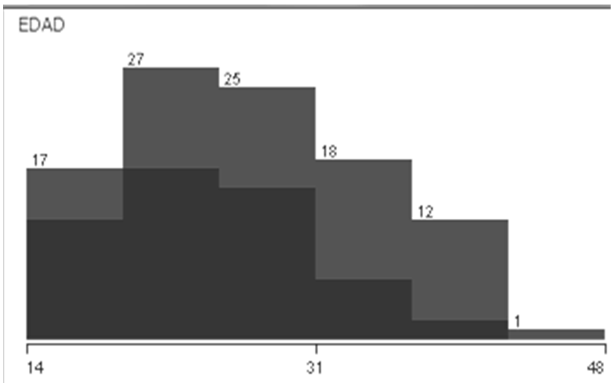


Gráfico 1: Histograma de la Edad de las pacientes de la muestra

Este histograma nos muestra el intervalo de edad de las pacientes de la muestra, la diferencia de colores determinan la clase a la que pertenece cada intervalo. Podemos observar lo siguiente:

- Las pacientes del intervalo 14-20 años pertenecen en su mayoría a la clase “sobreviviente”
- Las pacientes del intervalo 21-31 años tienen un mayor porcentaje de sobrevida, coincide con el promedio de edad adecuado y de la muestra.
- El intervalo de paciente entre 32-40 años tiene mayor porcentaje de muerte.
- Las pacientes mayores a 40 años pertenecen a la clase “fallecida” en gran porcentaje.

Resultados por algoritmo testeado:

Los algoritmos usados para evaluar la base de datos en mortalidad materna dieron como resultado cifras continuas indicando los siguientes sucesos: Tabla 2.

- **Especificidad:** es la probabilidad de que pacientes complicadas y de riesgo pertenezcan a la clase Sobreviviente. Es decir los verdaderos negativos.

$$Especificidad = \frac{VN}{VN + FP}$$

- Fracción de verdaderos negativos (FVN). Demuestra la cantidad de pacientes que realmente pertenecen a la clase Sobreviviente. Quiere decir que si el algoritmo estudiado tiene alto porcentaje de especificidad determina con gran éxito la probabilidad de sobrevida en pacientes complicadas durante su embarazo según los datos proporcionados en la ficha de antecedentes.

- **Clasificación correcta**: de la totalidad de datos, entre los que 52 que pertenecen a la clase Sobreviviente, y los 48 que pertenece a la clase Fallecida, determina dentro de cada clase cuantas instancias luego de la construcción del clasificador cuantas si pertenecen a la clase determinada.

- En el caso de pertenecer a la clase sobreviviente o a la fallecida de las 100 instancias cuantas fueron clasificadas correctamente.

- **Clasificación incorrecta:** del mismo modo la cantidad de instancias que no fueron clasificadas de manera correcta, son las que de manera supervisada se sabe que pertenecen a una u otra clase y fueron incluidas dentro de la cual no eran. Si el indicador emite un número mayor al 50% de la cantidad total de instancias, no se debe considerar como eficiente.

- **Sensibilidad:** es la capacidad del algoritmo de clasificar a las pacientes complicadas dentro de la clase Fallecidas. Es decir que si el clasificador tiene un alto porcentaje tiene mejor curva de corte y discernimiento entre los sujetos que pertenecen o no a la clase fallecida, es así que si la cifra de sensibilidad es del 90%, existe entonces esa probabilidad de que la paciente fallezca.

$$Sensibilidad = \frac{VP}{VP + FN}$$

- **Mean absolute error:** Se define **error absoluto** de una medida la diferencia entre el valor medio obtenido y el hallado en esa medida todo en valor absoluto.

- Entonces el promedio de error absoluto, es la suma de los errores absolutos de clasificación en cada uno de los sujetos llevados a promedio. El clasificador que arroje mayor cifra (mayor a 0.1) define un error de clasificación alto, por lo cual no se debe considerar por sobre los que arrojen una cifra menor.

- **Tiempo de ejecución:** medido en segundos es la cantidad de tiempo que demora en construir la arquitectura del clasificador y en arrojar resultados.
- Puede que un clasificador se defina como eficiente si el tiempo que emplea en emitir resultados es menor a 5 segundos, aun así depende de los demás indicadores para valerse de esta característica.

- **Kappa statistic:** el Kappa statistic es la concordancia de comparación que tienen los observadores de clasificación. Quiere decir en una matriz de clasificación, el índice esperado entre el diagonal principal esperada (Xii elemento clasificado en la misma clase por ambos observadores) y el índice real luego de la clasificación efectuada por la arquitectura seleccionada (sea regresión lineal, backpropagation, Naive-bayes, etc.), es la diferencia en porcentaje de su lejanía a este valor.

- Si por ejemplo, la matriz esperada clasifica el valor en 25.00 y el resultado de la arquitectura es 26.7, la diferencia seria, 1.7 equivale al 90.32%. Entonces cuanto mas grande sea el porcentaje, estará más cerca de ser considerado eficiente.

- **Root mean squared error:** error cuadrático medio, es una medida de uso frecuente de las diferencias entre los valores pronosticados por un modelo o un estimador y los valores realmente observados. RMSD es una buena medida de precisión, pero sólo para comparar diferentes errores de predicción dentro de un conjunto de datos y no entre los diferentes, ya que es dependiente de la una escala muestra. Estas diferencias individuales también se denominan residuos, y la RMSD sirve para agregarlos en una sola medida de la capacidad de predicción.

- **Relative absolute error:** es el error relativo a cada característica de la clase, por ejemplo el error relativo de tener de Espacio Intergenésico 0.5 años y pertenecer o no a la clase fallecida, la clasificación indicaría que si pertenece, por ser el valor indicado para aquellas pacientes que están en peligro.
  En este caso el valor positivo para pertenecer al clase sobreviviente es de entre 2-4 años o primera gesta: 0.5 años incluido en la clase fallecido si el error entre los valores determinados por la clase y el valor ingresado es menor a 1.

- **Root relative squared error:** La **raíz relativa** E **de error al cuadrado** $_i$ de un programa individual i es evaluado por la ecuación:

$$Ei = \sqrt{\sum_{j-1}^{n} \frac{\left(P_{(ij)} - T_j\right)^2}{\sum_{j-1}^{n}\left(T_j - \bar{T}\right)^2}}$$

donde $P_{(ij)}$ es el valor predicho por el programa para el individuo i j muestra de casos (de los casos de la muestra n), T_j es el valor objetivo para la muestra j caso, y $\bar{T}$ está dada por la fórmula:

$$\bar{T} = \frac{1}{n}\sum_{j-1}^{n} T_j$$

Para un ajuste perfecto, el numerador es igual a 0 y $E_i = 0$. Así, el E_i índice varía de 0 a infinito, con el ideal que corresponde a 0.

	Reglas	**Redes Bayesianas**	**Redes Neuronales**
INDICADORES	**OneR**	**NAÏVE BAYES KERNEL**	**RBFNetwork**
Especificidad	0.836	0.91	0.9
Clasificación correcta	84	91	90
Clasificación incorrecta	16	9	10
Sensibilidad	0.868	0.914	0.9
Mean absolute error	0.16	0.1142	0.1468

Tiempo de ejecución	0.2	0.01	0.26
Kappa statistic	0.6759	0.819	0.7997
Root mean squared error	0.4	0.2737	0.3032
Relative absolute error	32.03%	22.86%	29.39%
Root relative squared error	80.01%	54.74%	60.64%

Tabla III: Algoritmos de clasificación supervisada con mejores resultados por familia

INDICADORES	**Arboles de Decisión** **J48**	**Algoritmos de Distancia** **LWL**	**Regresión Logística** **Logistic**
Especificidad	0.9	0.889	0.82
Clasificación correcta	90	89	82
Clasificación incorrecta	10	11	18
Sensibilidad	0.9	0.902	0.82
Mean absolute error	0.1174	0.168	0.1753
Tiempo de ejecución	0.02	0	0.81
Kappa statistic	0.7994	0.6388	0.6388
Root mean squared error	0.299	0.3019	0.4157
Relative absolute error	23.75%	33.64%	35.10%
Root relative squared error	59.80%	60.39%	83.15%

Tabla IV: Algoritmos de clasificación supervisada con mejores resultados por familia

Descripción de la metodología propuesta

Se propone entonces que evaluando cada uno de los indicadores más importantes en la construcción del clasificador, se descarten aquellos que no cumplen las siguientes características:

- Especificidad > 90%
- Clasificación correcta > 90 instancias
- Clasificación incorrecta < 10 instancias
- Sensibilidad >90%
- Mean absolute error < 0.1 ideal
- Kappa statistic >0.79, >0.9 ideal
- Root mean squared error < 0.3, <1 ideal
- Relative absolute error <25%, <1 ideal
- Root relative squared <50%, 0 ideal

El clasificador que cumpla con estas especificaciones, se considera como optimo para la integración en un sistema que evaluará la base de datos que contengan los datos de las pacientes a comparar con un registro nuevo de entrada.

Etapa de Evaluación:

Cada uno de los clasificadores estudiados, que analizaron la muestra de pacientes, arrojaron los indicadores que muestra la tabla, en nivel de rendimiento y buenos resultados, aceptables para el estudio, la arquitectura que lleva los porcentajes mas óptimos dentro de los parámetros, es el algoritmo de Naive bayes con estimador de núcleo.

Descartando así el resto de arquitecturas como apropiadas para este tipo de muestras. Además de visualizar de manera más clara y correcta los resultados comunes entre cada una de las arquitecturas dejando atrás aquel paradigma que incluye a la regresión logística como la más adecuada a la hora de realizar diagnósticos preventivos en salud.

Kernel estimator, al ser un método de construcción no paramétrico, es mas flexible que los clasificadores que incluyen parámetros. Dividiendo la muestra en 40 grupos en la variable edad, paridad 15, controles 17, para la exploración descubre nuevas alternativas de clasificación en los atributos de la clase, mostrándolos como relevantes.

El problema más resaltante es que requiere mayor tamaño de muestra para mantener estos resultados, ya que utilizado el agrupamiento para la clasificación (clustering) que podría ser una desventaja si de clasificadores supervisados hablamos. Esto se denomina the curse of dimensionality, pues dependen de la elección de un parámetro suavizado, en este caso los valores la edad, número de controles, y número de hijos, transformándola en no objetiva.

Cuando Naive-Bayes actúa con la estimación no paramétrica, las estructuras que se construyen a partir de esta arquitectura se obtienen a partir de un árbol formado con variables potencialmente predictoras multidimensionales. (Serrano, 2010)

3 Conclusiones

Con respecto a la respuesta de las hipótesis podemos afirmar lo siguiente:

- Encontramos que las arquitecturas propuestas por esta memoria, totas conservan una efectividad bastante aceptable e cada uno de sus indicadores que en conjunto hacen un 80% de eficacia. Siendo el más optimo el de Naive-Bayer Kernel.

- Descarta estas hipótesis luego del trabajo realizado.

- Las Redes Bayesianas brindan al estudio de los datos en mortalidad y sobrevida materna una especificidad 91%, clasificación correcta 91%., error absoluto 0.1142, y sensibilidad 0.914 recomendada.

- Las Redes Neuronales brindan al estudio de los datos en mortalidad y sobrevida materna una especificidad 90%, clasificación correcta 90%, error absoluto 0.1468 y sensibilidad 90% recomendada.

- Los Arboles de Decisión brindan al estudio de los datos en mortalidad y sobrevida materna una especificidad 90%, clasificación correcta 90%, error absoluto 0.1174 y sensibilidad 90% recomendada.

- Los Algoritmos basados en Distancias brindan al estudio de los datos en mortalidad y sobrevida materna una especificidad 88.9%, clasificación correcta 89%, error absoluto 0.168 y sensibilidad 90.2% recomendada.

- La Regresión Logística brinda al estudio de los datos en mortalidad y sobrevida materna una especificidad 82%, clasificación correcta 82%, error absoluto 0.1753 y sensibilidad 82% recomendada.

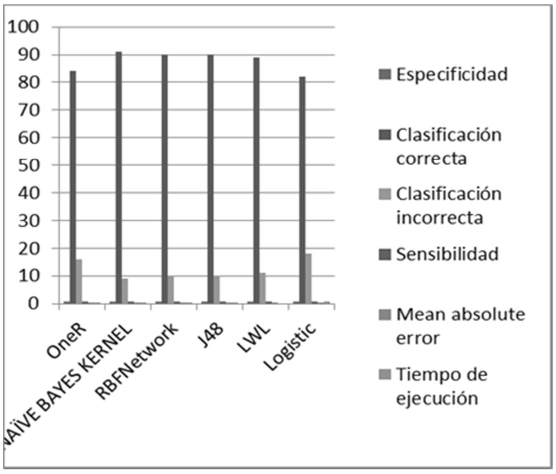


Gráfico 2: Indicadores estadísticos de todos los algoritmos de clasificación supervisada

4 Recomendaciones

- Se recomienda usar datos discretizados, si se desea usar Naive-Bayes para su evaluación.
- Agrupar los datos de la manera propuesta en la Tabla 1, así podremos respetar los parámetros acerca de mortalidad materna que establecen los expertos salubristas.
- Es importante comparar los resultados con la regla de decisión OneR para próximos experimentos, pues nos da la mejor noción de veracidad y efectividad a la hora de analizar la información.
- Se pretende implementar un sistema en R Project, para ingresar nuevos registros y que lleve en la memoria la base de datos recolectada a través de este trabajo.
- Para ayudar a la muestra numeraria que requiere Naive- Bayes Kernel, es necesario ingresar los registros de las muertes maternas y complicaciones diarias de las gestantes a nivel Nacional en la base de datos.

Agradecimientos

Agradezco al Dr. Basilio Araujo y al Dr. Yosu Yuramendi, por su apoyo y perseverancia para la culminación de este proyecto de fin de master.

Al personal de salud que proporciono la información clínica que se usó en la investigación.

5 Síntesis curricular de la autora:

Pilar Hidalgo León, Ingeniera de Sistemas de la Universidad Andina del Cusco en Perú, Docente en la Facultad de Ingeniería, Sustentación de proyecto de máster en la Universidad del País Vasco. *Contacto:*phidalgo@uandina.edu.pe, Universidad Andina de Cusco, San Jerónimo, Cusco, Perú.

6 Referencias bibliográficas

Clasificadores supervisados: el objetivo es obtener un modelo clasificatorio valido para permitir tratar casos futuros.(2006) Araujo, B. S. Aprendizaje Automatico: conceptos básicos y avanzados. Madrid, España: Pearson Prentice Hall,

Cordero Muñoz, L., Luna Flórez, A., & Vattuone Ramírez, M. (2010) Salud de la mujer indígena : intervenciones para reducir la muerte materna. © Banco Interamericano de Desarrollo.

Antonio Serrano, E. S., (2010) Redes Neuronales Artificiales. Universidad de Valencia.

Calderón Saldaña, J., & Alzamora de los Godos Urcia, L. (2009) Regresión Logística Aplicada A La Epidemiología. Revista Salud, Sexualidad y Sociedad, 1[4].

Calderón, S. G, (2006) Una Metodología Unificada para la Evaluación de Algoritmos de Clasificación tanto Supervisados como No-Supervisados. México D. F.: resumen de tesis doctoral.

Corso, C. L. (2009) Aplicación de algoritmos de clasificación supervisada usando Weka. Córdoba: Universidad Tecnológica Nacional, Facultad Regional Córdoba.

María García Jiménez, & Aránzazu Álvarez Sierra. (2010) Análisis de Datos en WEKA – Pruebas de Selecitividad. Articulo.

María N. Moreno García* L. A. (2005) Obtención Y Validación De Modelos De Estimación De software Mediante Técnicas De Minería De Datos. Revista colombiana de computación, 3[1], 53-71.

Msp Mynor Gudiel M., 1. E. (2001-2002) Modelo Predictor De Mortalidad Materna. Mortalidad Materna, 22-29.

Organización Mundial De La Salud. Mortalidad Materna En 2005, (2008) : Estimaciones Elaboradas Por La Oms, El Unicef, El Unfpa Y El Banco. Ginebra 27: Ediciones de la OMS.

Porcada, V. R. (2003) Clasificación Supervisada Basada En Redes Bayesianas. Aplicación En Biología Computacional. Madrid: Tesis Doctoral.

Ramírez Ramírez, R., & Reyes Moyano, J. , (2011). Indicadores De Resultado Identificados En Los Programas Estratégicos. Lima: Encuesta Demográfica Y De Salud Familiar – Endes

Ramirez, C. J. (2009)Caracterización De Algunas Técnicas Algoritmicas De La Inteligencia Artificial Para El Descubrimiento De Asociaciones Entre Variables Y Su Aplicación En Un Caso De Investigación Específico. Medellin: Tesis Magistral.

Reproductive Health Matters, (2009) Mortalidad Y Morbilidad Materna:Gestación Más Segura Para Las Mujeres. Lima: © Reproductive Health Matters.

Vilca, C. P. (2009) Clasificación De Tumores De Mama Usando Métodos. Lima: Tesis.

Dirección Regional De Salud Cusco, (2007) Análisis De Situación De La Mortalidad Materna Y Perinatal, Región Cusco.

Ministerio De Salud, Dirección General De Epidemiología Situación De Muerte Materna, (2010-2011), Perú.