

Knowledge Discovery in Online Repositories: A Text Mining Approach

Fatudimu I.T

*Department of Computer and Information Science
Covenant University, Ota, Nigeria
E-mail: ibkfat@yahoo.co.uk
Tel: +234-08052318494*

Musa A.G

*Department of Computer and Information Science
Covenant University, Ota, Nigeria
E-mail: adebola.musa@covenantuniversity.com*

Ayo C.K

*Department of Computer and Information Science
Covenant University, Ota, Nigeria
E-mail: ckayome@yahoo.com*

Sofoluwe A. B

*Department of Computer Science, University of Lagos
Lagos, Nigeria
Email: absofoluwe@yahoo.com*

Abstract

Before the advent of the Internet, the newspapers were the prominent instrument of mobilization for independence and political struggles. Since independence in Nigeria, the political class has adopted newspapers as a medium of Political Competition and Communication. Consequently, most political information exists in unstructured form and hence the need to tap into it using text mining algorithm.

This paper implements a text mining algorithm on some unstructured data format in some newspapers. The algorithm involves the following natural language processing techniques: tokenization, text filtering and refinement. As a follow-up to the natural language techniques, association rule mining technique of data mining is used to extract knowledge using the Modified Generating Association Rules based on Weighting scheme (GARW).

The main contributions of the technique are that it integrates information retrieval scheme (Term Frequency Inverse Document Frequency) (for keyword/feature selection that automatically selects the most discriminative keywords for use in association rules generation) with Data Mining technique for association rules discovery. The program is applied to Pre-Election information gotten from the website of the Nigerian Guardian newspaper. The extracted association rules contained important features and described the informative news included in the documents collection when related to the concluded 2007

presidential election. The system presented useful information that could help sanitize the polity as well as protect the nascent democracy.

Keywords: Text Mining, Data Mining, Association Rule Mining, Inference, Politics

1.0 Introduction

We have entered an era where very large amount of politically oriented text are now available online. This includes both official documents, such as the full text of laws and the proceedings of legislative bodies, and unofficial documents, such as postings on weblogs (blogs) devoted to politics [1].

Fortunately, there are many tools at our disposal to manage this outbreak of textual information, many of these tools are derived from earlier works in Information Retrieval (IR), Natural language processing, and statistics, Artificial intelligence (AI), Information Theory and Data Mining [2].

Social scientists often analyze textual data for indicators about the source, purpose, and consequences of communications. In media and political analyses, for instance, texts are scrutinized for evidence of thematic trends and framing, or the packaging of information with the intent of creating a particular interpretation [3, 4].

The term text mining was coined to describe tools used to manage textual information. Text mining, defined as knowledge discovery in textual databases [5], allows us to create a technology that combines a human's linguistic capabilities with the speed and accuracy of a computer. Text mining aims at employing technology to analyze more detailed information in the content of each document and to extract interesting information that can be provided only by multiple documents viewed as whole, such as trends and significant features that may be a trigger to useful actions and decision-making [8]. However, data mining is the analytical process designed to explore structured data in search of consistent patterns and /or systemic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data [9].

Nigeria as the most populous nation in Africa, with an estimated 132 million people, had faced intermittent political turmoil and economic crisis since gaining independence in October 1960. Nigerian political life has been scarred by conflict along both ethnic and geographic lines and misrule has undermined the authority and legitimacy of the state apparatus [6].

After 16 years of military rule, Nigeria made a transition to civilian governance in 1999. In May 2007, after two terms in office, power was handed over to another president elect. The general election was significant because it marked the country's first transfer of power from one civilian government to another. Some analysts suggested that the election would result in a threat to domestic regional tensions triggering wider civil unrest. They also suggested that controversies surrounding the elections themselves could lead to an electoral delay or violence. Therefore, credible and peaceful elections are critical to the future of both the country and the region [6].

Even though, this all important election had come and gone, it is important to have a system in place that would have predicted what was experienced during this election and therefore be able to better advice the necessary authorities as to how to prepare for it.

This paper therefore is aimed at implementing a text mining algorithm on some data available in the Nigerian Guardian website in order to generate useful political inferences. In making decisions concerning the future of elections, the authorities concerned are much more likely to consult rules generated from such a system in order to know what the outcome of the elections would look like rather than reading through the heap of online newspapers one after the other.

The rest of the paper is organized as follows. Section 2.0 presents a review of related work, section 3.0 presents the objective, 4.0 describes the methodology used and the system architecture is presented in section 5.0. Experiments, interpretation and discussion are presented in section 6.0. Section 7.0 provides conclusion and future work.

2.0 Previous Research

In their work on semantic annotation, Mingcai Hong et.al. [10], addressed the issue of semantic annotation using horizontal and vertical contexts. OnTeA: (Semi-automatic Ontology based Text Annotation Method) was developed by Michal Laclavík et.al. [11]. It describes a solution for the ontology-based text annotation tool to analyze a document or text using regular expression patterns and detects equivalent semantic elements according to the defined domain ontology. Fabio Ciravegna et al [12] developed a tool, Melita, for the definition and development of ontology-based annotation services which does beyond the dichotomy rule learning versus rule writing of classic annotation systems, as it allows adopting different strategies, from annotating examples in a corpus for training a learner to rule writing and even a mixture of them.

In addition to this, there are a number of annotation tools and approaches such as CREAM [11, 14] or Magpie [11, 13] which provide users with useful visual tools for manual annotation, web page navigation, reading semantic tags and browsing [11,15] or providing infrastructure and protocols for manual stamping of documents with semantic tags. The basic limitation of these systems is that they are geared towards information extraction and not knowledge discovery. As mentioned earlier, the result of an annotated text is used for further computer processing, for example, using semantic data in knowledge management [11, 13] or in Semantic Organization applications. This further processing for example can be captured in Text mining.

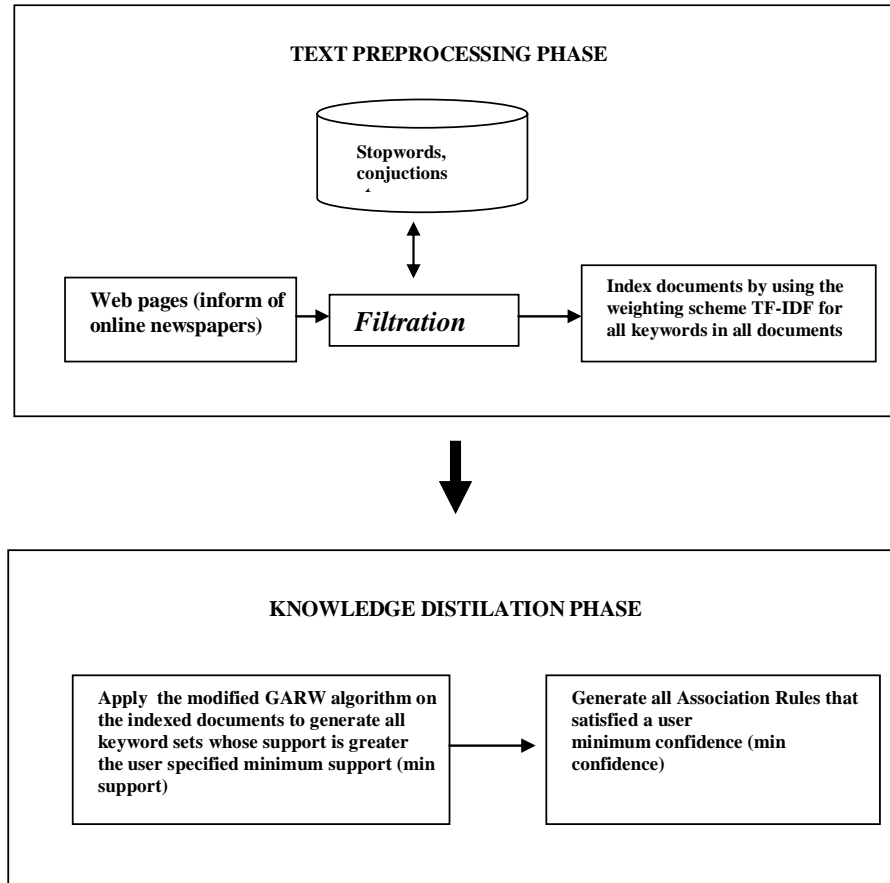
Ronen Feldman et.al. [16], developed a Document Explorer, a tool that implements text mining at the term level. Earlier works on mining association rules from text have explored the use of manually assigned keywords. Where, they used keywords as features for generation of association rules [17, 18]. The drawbacks of approaches that use manually assigned keywords are that: (1) it is time consuming to manually assign the keywords; (2) the keywords are fixed (i.e., they do not change over time or vary based on a particular user); (3) if the keywords are manually assigned, they are subject to discrepancy; (4) the textual resources are constrained to only those that have keywords. Several other researchers [19,20,21] applied existing data mining techniques to discover episode rules from texts, where Episode rule mining is used for language analysis because it preserves the sequential structure of terms in a text document. However, in our work we focus on the extraction of association rules that present the relations existing among the keywords in texts ignoring the order in which the keywords occur.

3.0 Objective of Research

The objective of this paper is to develop a text mining system based on the modified GARW algorithm that will generate inferences from political dataset.

4.0 Research Methodology

The methodology used for this research is based on the modified architecture in figure 1.

Figure 1: Modified Text Mining System Architecture

The main difference between the architecture and the other existing ones is the inclusion of the component in text preprocessing phase that is responsible for indexing of documents based on the TF-IDF weighing scheme.

Basically, the above architecture is divided into two phases:

1. **The text preprocessing phase:** This phase is aimed at optimizing the performance of the next phase. It consists of text filtration, indexing and refinement of the extracted keywords based on the weighing scheme (TF-IDF).
 - a. **Filtration:** A word is selected as a keyword if it does not appear in a pre-defined stop-words list. The stop-words list consists of articles, pronouns, determinants, prepositions and conjunctions, common adverbs and non-informative verbs.
 - b. **Indexing and refinement:** The techniques for automated production of indexes associated with documents usually rely on frequency-based weighing schemes. The weighing scheme TF-IDF (Term Frequency, Inverse Document Frequency) is used to assign higher weights to distinguished terms in a document, and it is the most widely used weighing scheme which is defined as [7,22]:
 - $w(i,j)$ is known as the weighing scheme and could be greater than 0.
 - Nd_{i,t_j} is the number of times the term t_j occurs in the document d_i .
 - Nt_j is the number of documents in the collection C in which the term t_j occurs at least once.
 - $|C|$ is the number of documents in the collection C .

In general, this weighing scheme includes the intuitive presumption that the more often a term occurs in a document, the more it is representative of the document (term frequency) and the more the documents the term occurs in, the less discriminating it is (inverse document frequency). The system

sorts the keywords based on their scores and selects them based on the given weight chosen as threshold.

2. **The Knowledge distillation phase:** Knowledge is distilled using the GARW (Generating Association Rules based on Weighting scheme) algorithm described below:

Generating Association Rules Based on Weighting Scheme (GARW) algorithm

Given a set of terms

$$A = \{w_1, w_2, \dots, w_n\}$$

A Set of indexed documents

$$D = \{d_1, d_2, \dots, d_n\}$$

- d_1, \dots, d_n are indexed documents that contains keywords.
- Those keywords are also members of A i.e. the general database of keywords.

Association Rule

Association rule is one of the most important techniques in Data Mining. The problem of association rule mining deals with how to discover association rules that have support and confidence greater than the user-specified minimum support and minimum confidence. It is intended to capture dependency among items in the database.

The support of an item set is the fraction of transactions in the database that contain all the items in the database

$$Support(w_i w_j) = \frac{\text{support count of } W_i, W_j}{\text{Total number of documents}}$$

The confidence of rule a (association rule) $W_i \rightarrow W_j$ can be defined as the proportion of those transactions containing W_i that also contain W_j .

$$Confidence(W_i/W_j) = \frac{\text{Support } W_i W_j}{\text{Support } (W_i)}$$

The algorithm for generating association rules based on the weighting scheme is given as follows:

1. Scan the file that contains all the keywords that satisfy the threshold weight value and their frequency in each document.
2. Let N denote the number of top keywords that satisfy the threshold weight value.
3. Store the top N keywords in index file along with their frequencies in all documents, their weight values TF-IDF and documents ID in the following format: <doc-id><keyword><keyword frequency><TF-IDF>
4. Scan the indexed file and find all keywords that satisfy the threshold minimum support. These keywords are called large frequency1-keywordSet L_1 .
5. When K is greater than 2, (Note K is a keyword set having k-keywords sets). The candidate keywords C_k of size K are generated from large frequent (k-1) keywords sets, L_{k-1} that is generated in the last step.
6. Scan the index file, and compute the frequency of candidate keyword sets C_k that is generated in step 4.
7. Compare the frequencies of candidate keywords sets with minimum support.
8. Large frequent keyword sets L_k , which satisfy the minimum in support, is found from step 7 above.
9. For each frequent keyword set, find all the association that satisfies the threshold minimum confidence.

Rule post processing

We refined the rules generated by using parameters such as the support and confidence which in this case has already been included in the GARW algorithms above. One particular aspect of rule mining in text is that often a high support means the rule is too obvious and thus less interesting. Another technique that was used to remove unwanted rules is to specify stop rules i.e. rules that are common and can be removed automatically [7]. Association rules are easy to understand and to interpret for an analyst or a normal user. However, it should be mentioned that the association rule extraction is of exponential growth and a very large number of rules can be produced

5.0 Experiments, Interpretation and Discussion

A. Data description

In order to extract association rules from texts, we applied a selected sample of 100 recent WebPages news that are related to the politics and policies in Nigeria in the period from 1 January 2007 to 20 April 2007. The source for this news is The Guardian News paper website. The collection of the 100 documents (corpus) is 1.05 MB in size and contained 249100 single words. Each document contained on average 2491 single words. After the filtration process, the collection of documents contained 240099 single words. The system is implemented using C language and executed on a Pentium 4, 2.2 GHz system running Windows XP professional with 512 MB of RAM.

B. Description of the extracted association rules

- Finding association rules in text documents can be useful in a number of contexts such as investigation and general understanding of events in real world. The text in figure 2 is a segment of an update of the election status in Nigeria from the guardian Newspaper website already converted to textual document.

Figure 2: An excerpt from part of the news documents.

Friday, March 30, 2007

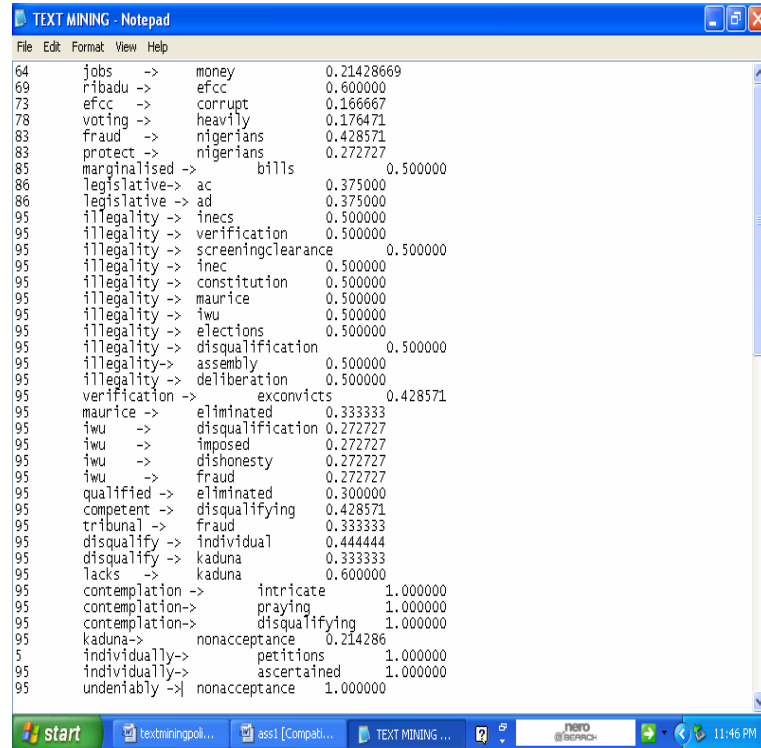
Adebayo Adefarati: The man and his unfulfilled dreams
By Clifford Ndujihe Abiodun Fanoro and Idowu Ajanaku (Lagos) and Iyabo Lawal (Ibadan)

WHEN the late Chief Adebayo Adefarati joined the presidential race recently on the plank of the Alliance for Democracy (AD), he shocked some observers, giving his support for South-South presidency.

During the 2005 political reforms conference, he urged delegates to hammer out modalities to ensure that the zone produced the next president because of years of marginalisation of the area, which produces crude oil, the mainstay of the nation's economy.

The late politician however, had a special reason for joining the race. Apart from service to humanity, which he said influenced into politics, Adefarati did not want the mistake the party and Afenifere made in 2003 when they had an "unholy" alliance with President Olusegun Obasanjo, which led to the defeat of the AD in the South-West.

In this document, there are multiple features scattered widely. The aim of this work is to find relations between features and present them in association rules form, to give the user or reader of the news the useful information about the Nigerian 2007 elections. In extracting these rules, we do not take into account the order in which the word occurs but extracted useful information from the document based on abstractions that describe the relationships between the features in the texts.

Figure 3: Part of the resultant association rules generated from the mined documents


Document ID	Term	Right-hand Side Term	Confidence
64	jobs	money	0.21428669
69	ribadu	efcc	0.600000
73	efcc	corrupt	0.166667
78	voting	heavily	0.176471
83	fraud	nigerians	0.428571
83	protect	nigerians	0.272727
85	marginalised	bills	0.500000
86	legislative	ac	0.375000
86	legislative	ad	0.375000
95	illegality	inecs	0.500000
95	illegality	verification	0.500000
95	illegality	screeningclearance	0.500000
95	illegality	inec	0.500000
95	illegality	constitution	0.500000
95	illegality	maurice	0.500000
95	illegality	iwu	0.500000
95	illegality	elections	0.500000
95	illegality	disqualification	0.500000
95	illegality	assembly	0.500000
95	illegality	deliberation	0.500000
95	verification	exconvicts	0.428571
95	maurice	eliminated	0.333333
95	iwu	disqualification	0.272727
95	iwu	imposed	0.272727
95	iwu	dishonesty	0.272727
95	iwu	fraud	0.272727
95	qualified	eliminated	0.300000
95	competent	disqualifying	0.428571
95	tribunal	fraud	0.333333
95	disqualify	individual	0.444444
95	disqualify	kaduna	0.333333
95	lacks	kaduna	0.600000
95	contemplation	intricate	1.000000
95	contemplation	praying	1.000000
95	contemplation	disqualifying	1.000000
95	kaduna	nonacceptance	0.214286
95	individually	petitions	1.000000
95	individually	ascertained	1.000000
95	undeniably	nonacceptance	1.000000

The Argumentation of the Thresholds Chosen

In text mining in general, a very large number of association rules are found. So the measures like support and confidence are important when creating keyword sets and selecting the final rules. However, the problem is that we may find the important keywords which have frequently appeared recently but not discovered because the height of support and confidence threshold values. In order to have a fair representation of the important keywords in the corpus to be mined, we selected a TF-IDF threshold of 5.0. This helped us to find informative keywords to extract rules from. Furthermore, a low threshold support of 2% was used so as to extract important keywords (such as dishonesty, iwu) that would not have appeared if we chose high support value, and these keywords happen to be very informative regarding Nigerian politics. Lastly, we chose higher threshold confidence value 15% to make sure that the final rules gotten from the system are the most interesting ones.

Interpretation of the extracted rules

Some of the association rules abstractions that describe the relations between features in the texts are presented in figure 3 above. The rules give information about the possible occurrence during and after the elections. The first column represents the Document identification which reveals the identification number of the document from which the rule originated. The second column is the generated association rules and the third column is the confidence of the rules generated. Samples of abstractions and their corresponding extracted association rules are shown below.

Table 1: Samples of extracted association rules

Document ID	Association rules	Confidence Threshold(/100)
83	protect --> Nigerians	0.272727
95	illegality --> elections	0.500000
95	iwu --> dishonesty	0.272727
95	competent --> disqualifying	0.428571
6	rally --> deceit	0.230760
6	difficult --> roads	0.280000

We observe from the output of our system, the following sample features and our system presents the relationships between them: protect, nigerians, illegality, elections, iwu, dishonesty, competent, disqualifying, rally, deceit, difficult, roads. The rule: *protect--> nigerians* gives a clear inference that Nigerians need to be protected during the election and this can be validated by the occurrences of violence that were observed during the election. The rule: *illegality-> elections* would help to warn the body in charge of organizing the elections that there would probably be sorts of illegality perpetrated during the election. The rule: *rally -> deceit* can help to infer lack of sincerity attached to the rallies that were organized for the election earring campaigns. The rule: *iwu-> dishonesty* would help to keep the populace on the alert because the chairman of the electoral body (iwu) might indulge in some dishonest practices. etc

7.0 Conclusion and Future Work.

Findings from the system reveal that there is strong relationship between the opinions of stakeholders before the 2007 presidential election in Nigeria and the aftermath of the election. This paper has presented a text mining technique for automatically extracting association rules from a collection of documents based on the keyword features. The system is domain independent so it is deplorable on different domains. The system can be applied on all or specific parts of documents. In addition, it is designed to automatically index documents by labeling each document by a set of keywords that satisfy the given weight constraints based on the weighting scheme. For future research, we plan to convert documents to XML before processing, so as to be able to generate index that better describes extracted features. Generally, the issues generated by the system are the same issues that are being challenged by the stakeholders in the court of law, even one year after the conduct of the election. Therefore, this system could be employed by government to help predict future trends of elections with a view to safeguarding the nascent democracy.

References

- [1] Matt Thomas, Bo Pang, and Lillian Lee: Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. 2006.
- [2] Manu Konchady : Text Mining Application Programming. Pp 2-30, 2006
- [3] Karin Verspoor, Antonio Sanfilippo, Mark Elmore and Ed MacKerrow: Deploying Natural Language Processing for Social Science Analysis: 2004
- [4] Goffman, Erving: Frame Analysis: An Essay on the Organization of Experience. London: Harper and Row, 1974
- [5] Ah-Hwee Tan: Text mining: The state of the art and the challenges, 2006.
- [6] Lauren Ploch: CRS Report for Congress: Nigeria: Current Issues: Congressional Research services, Order Code RL33964, Pp CRS 1-23, 2007.
- [7] Hany Mahgoub; Dietmar Rosner; Nabil Ismail; Fawzy Torkey . A Text Mining Technique Using Association Rules Extraction, INTERNATIONAL JOURNAL OF COMPUTATIONAL INTELLIGENCE VOLUME 4 NUMBER 1 2007 ISSN 1304-2386.
- [8] Nasukawa T. , Nagano T., Text Analysis and Knowledge Mining System, 2006
- [9] <http://www.statsoft.com/textbook/stdatmin.html#mining>.
- [10] Mingcai Hong, Jie Tang, and Juanzi Li: Semantic Annotation using Horizontal and Vertical Contexts, 2004.
- [11] Michal Laclavik¹, Martin Seleng¹, Emil Gatial¹, Zoltan Balogh¹, Ladislav Hluchy¹: Ontology based Text Annotation, 2004
- [12] Fabio Ciravegna, Alexiei Dingli, Jose' Iria, and Yorick Wilks: Multi-strategy Definition of Annotation Services in Melita, 2002
- [13] Domingue J., Dzbor M.: Magpie: supporting browsing and navigation on the semantic web. In IUI '04, pages 191-197, New York, NY, USA, 2004. ACM Press. ISBN 1-58113-815-6.
- [14] Handschuh S., Staab S.: Authoring and annotation of web pages in cream. In WWW '02, pages 462-473, NY, USA, 2002. ACM Press. ISBN 1-58113-449-5. doi: <http://doi.acm.org/10.1145/511446.511506>.
- [15] Uren V. et al.: Browsing for information by highlighting automatically generated annotations: a user study and evaluation. In K-CAP '05, pages 75-82, NY, USA, 2005b. ACM Press. ISBN 1-59593-163-5
- [16] Ronen Feldman, Moshe Fresko, Haym Hirsh, Yonatan Aumann,* Orly Liphstat, Yonatan Schler, Martin Rajman: Knowledge Management: A Text Mining Approach: Proc. of the 2nd Int. Conf. on Practical Aspects of Knowledge Management (PAKM98) Basel, Switzerland, 29-30 Oct. 1998, (U. Reimer, ed.)
- [17] Feldman R. and Dagan I., "Knowledge discovery in textual databases (KDT)", in *Proc. 1st Int. Conf. on Knowledge Discovery and Data Mining*, 1995.
- [18] Feldman R. and Hirsh H., "Mining associations in text in the presence of background knowledge," in *Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining*, Portland, USA, 1996.
- [19] Ahonen H., Heinonen O., Klemettinen M., and Inkeri V., "Mining in the phrasal frontier," in *Proc. PKDD'97.1st European Symposium on Principle of data Mining and Knowledge Discovery*, Norway, June, Trondheim, 1997.
- [20] Ahonen H., Heinonen O., Klemettinen M., and Inkeri V., "Applying data mining technique for descriptive phrase extraction in digital document collections" in *Proc. of IEEE Forum on Research and technology Advances in Digital Libraries*, Santa Barbra CA, 1998.
- [21] Mannila H., Toivonen H. and Verkamo A. I., "Discovery of frequent episodes in event sequences," *Data Mining and Knowledge Discovery*, 1(3), 1997b, pp. 259-289.
- [22] Rajman M. and Besancon R., "Text mining: natural language techniques and text mining applications", in *Proc. 7th working conf. on database semantics (DS-7)*, Chapan & Hall IFIP Proc. Series. Leysin, Switzerland Oct. 1997, 7-10.

- [23] Uren V. et al.: Semantic annotation for knowledge management: Requirements and a survey of the state of the art. *Journal of Web Semantics: Science, Services and Agents on the WWW*, 4(1):14-28, 2005.