

Applying an Enhanced Algorithm for Mining Incremental Updates on an Egyptian Newspaper Website

Laila Mohamed ElFangary
Information Systems Department
Faculty of Computers and Information - Helwan University
Cairo, Egypt
e-mail: lailaelfangary@gmail.com

Abstract—This paper presents new approaches for analyzing Egyptian Newspaper websites and presenting the most interesting topics that attract the largest portion of reading population to their newspaper website. The paper also presents techniques used for text and news mining. We first start with topic detection and tracking techniques to collect the reader website navigation data for the various topics then we use the proposed incremental algorithm (PIA), our previously published algorithm that was implemented in our previous researches, for mining of the news database. The paper concludes with significant results for identifying the readers trends, thus can dynamically display the most expected and required headline topics to the reader. The paper discovers also the most important topics for the users. Moreover, the paper had a significant effect for guiding the newspaper for the appropriate method for modifying its website to be dynamic according to the user interests specified by the association rules algorithm. We believe that this work had high impact in increasing the number of visitors for the newspaper website and attracting the reader's attention.

I. INTRODUCTION

Egyptian Newspapers websites which concentrate on their traditional ways for displaying headlines are faced with a decreasing readership. This is due to the difficulties faced by the Journalists to present the most interesting topics that attract the largest portion of reading population to their newspaper website. This paper presents techniques used for text and news mining. An Egyptian newspaper website is mined using an enhanced algorithm for mining incremental updates in large databases. The process starts with topic detection and tracking technique for the visitor's website navigations for the various news topics, then applying the association rules algorithm on the extracted data to dynamically display the most expected and required headline topics to the reader. Moreover, the proposed algorithm analyzes the reader's opinions for the topics presented by the website. Mining the web site revealed the order of the most visited categories. Furthermore, the most frequent issues of concern to the web site visitors are related to every day cost of living including food and transportation. The previously mentioned results conclude that the site should be dynamically changing according to the user interests specified by the association rules algorithm. This will help in increasing the number of visitors and attracting the reader's attention.

II. BASIC TERMINOLOGY

A. Text Mining

Text mining involves extracting information from text that is useful for particular purposes through inferring structure from sequences representing natural language text. Adaptive techniques in text compression were applied in text mining including extraction of hierarchical phrase structures from text, identifying of key phrases in documents, locating proper names and quantities of interest in a piece of text, text categorization, word segmentation, acronym extraction, and structure recognition. [2] Moreover, documents are labeled by keywords, and knowledge discovery is performed by analyzing the co-occurrence frequencies of the various keywords labeling the documents. Keyword-frequency approach support a range of Knowledge data discovery operations, providing a suitable foundation for knowledge discovery and exploration for collections of unstructured text. [1] Data may be parsed into vectors of terms to extract information from free form text data. Information extracted from the vectored data can be used in applying methods of clustering for finding patterns in unstructured text information. Studies show how feature variables can be created from unstructured text information and used for prediction. [4]

B. News Mining

Text mining techniques that uncover trends discover associations and detect deviations from news notes were used to explore the society interests. Statistical representation of news reports using frequencies and probability distributions of topics and statistical measures using average or median, standard deviation and correlation coefficient were applied for analysis and discovery of useful information. [3] A VoiceTone Daily News data mining tool was used for analyzing spoken dialog systems in call centers. Relevant business and dialog features were extracted from the speech logs of caller-system interactions and tracked by a trend analysis algorithm. Alerts on multiple data streams were generated avoiding redundant "knock-on" alerts. [5]

In order to improve on complete-page mining of Web newspapers individual news items from the web pages are extracted and mined separately. Strategies using pattern-detection facilitating automatic news item extraction and results from clustering the extracted news items were presented. [7] News Web sites were mined by Hyperlink Induced Topics Search (HITS) algorithm, using an entropy-based analysis (LAMIS) mechanism for analyzing the entropy of anchor texts and links to eliminate the redundancy of the hyperlinked structure so that the complex structure of a Web site can be distilled. To eliminate redundancy an nfoDiscoverer mechanism was proposed. It applies the distilled structure to identify sets of article pages and employs the entropy information to analyze the information measures of article sets and to extract informative content blocks from these sets. Precision and the recall of those approaches were much superior to those obtained by conventional methods in mining the informative structures of news Web sites. [6] A learned information extraction system called DISCOTEX (Discovery from Text Extraction) was developed to transform text into more structured data which is then mined for interesting relationships. Rules mined from a database extracted from a quantity of texts are used to predict additional information to extract from future documents, thereby improving the recall of the underlying extraction system. These techniques were successfully applied to several computer job announcement postings from an Internet newsgroup. [8]

Text mining algorithms applied on newspapers involved natural language processing techniques as tokenization, text filtering and refinement. Association rule mining techniques of data mining were used to extract knowledge using a Modified Generating Association Rules based on Weighting scheme (GARW) that integrated information retrieval scheme with Data Mining technique for association rules discovery. [10] Incremental document clustering algorithms are used to address the special requirements in news clustering such as high rate of document update, or ability to identify event level clusters as well as topic level clusters. Topically related concepts and terms that are not explicitly linked, and topic ontology characterize news topics at multiple levels of abstraction for rich semantic information retrieval within Web news services. [9]

Only three out of the top twenty most visited Egyptian newspapers web sites enabled visitors to discuss their opinion on news topics. An enhanced algorithm for mining incremental updates is applied on one of those three Egyptian news paper websites. The website had the feature of taking the visitors votes (opinion) on the topics presented on the news journal.

C. Proposed Incremental algorithm (PIA)

Maintenance of large itemsets has been an important issue. When the existing database is updated by adding new transactions or deleting existing ones (Dynamic Databases), how can we update the association rules already discovered in the set of old transactions efficiently?. Naturally, when new transactions are added to a database, some of the existing frequent patterns may disappear whereas new frequent patterns that did not exist before may also be emerged.

The straight forward solution is to rerun the normal algorithms on the whole transaction database, i.e., the old database plus the new transactions. However, this process is not efficient because it ignores the previously discovered rules. Such approaches can waste a lot of computational and I/O resources, and result in relatively slow response times.

In our previous researches we implemented an association rule algorithm to address this problem using what we called the proposed incremental algorithm (PIA) [5] and we considered this problem within the context of association rule mining, a key data mining task.

The algorithm is designed to efficiently update large itemsets by taking set of previously discovered rules into account using some hypothesis to remove some of the old large itemsets or to add new large itemsets without doing much work. The algorithm had proved that it leads into significant I/O and computational savings and relatively faster response time in dynamic databases rather than rediscovering all the patterns by scanning the entire old and new databases.

III. DATA COLLECTION AND PREPROCESSING

Once the objective for the KDD (Knowledge Discovery in Database) process is known, a target data set must be assembled. As data mining can only uncover patterns already present in the data, the target dataset must be large enough to contain these patterns.

A. Data Collection:

Data was collected from an episode database for one of the top newspapers in Egypt. The news paper has a website that is used to present some news in different topics, for instance politics, athletic, etc. the website records some useful data in a database. These data are related to news paper visitors and their navigation through the website news, moreover the database contains their opinions about different topics.

Data collection process in data mining and knowledge discovery is necessary as it ensures that data gathered is both defined and accurate and that subsequent decisions based on arguments in the findings are valid.

B. Data Processing for the news database:

Because it was the policy of the newspaper to keep a log for all the useful information, we have full records about the visitors data and their IPs and their navigation for the various topics.

The Information Service Department of the newspaper presented different files containing the previously mentioned data. This data was processed by the algorithm to create a combined file in a transactional format with two columns, the transaction ID and the topics visited by the visitor; this enabled the discovery of associations between kinds of topics.

C. Applying Association Rules Technique on the database:

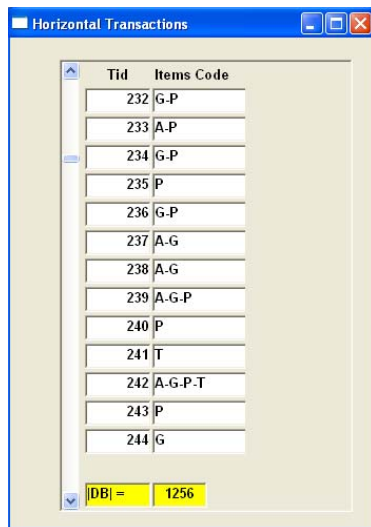
The Proposed Incremental algorithm (PIA) generated many rules, some of them were interested to the domain experts others were not, the most interesting rules for the newspaper staff was the dependence of certain topics on other topics.

The rules discovered were then made available to the staff for comment. These rules have been approved by the journalists and specialists.

IV. APPLYING INCREMENTAL ASSOCIATION RULES ALGORITHMS ON THE NEWS DATA

This section contains snapshots for some of the screens for the Proposed Incremental algorithm (PIA)

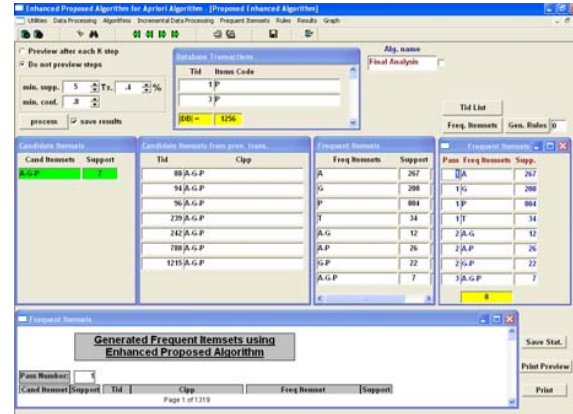
A. Snapshot of the algorithm implementation:



Tid	Items Code
232	G-P
233	A-P
234	G-P
235	P
236	G-P
237	A-G
238	A-G
239	A-G-P
240	P
241	T
242	A-G-P-T
243	P
244	G

DBI = 1256

Figure (1) Transactional database after conversion into the required association rules form

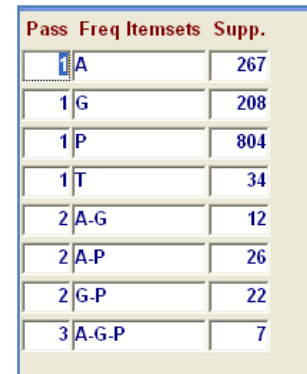


min. supp. 5 min. conf. 80% DBI = 1256

Generated Frequent Itemsets using Enhanced Proposed Algorithm

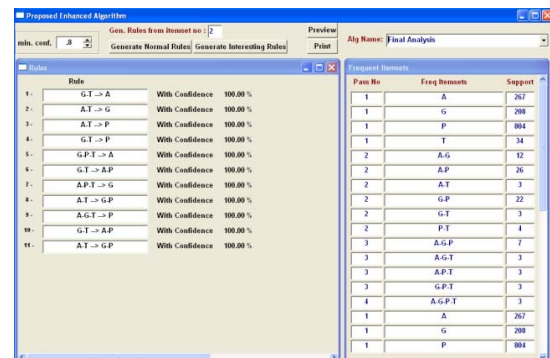
Pass	Freq Itemsets	Support
1	A	267
1	G	208
1	P	804
1	T	34
2	A-G	12
2	A-P	26
2	G-P	22
3	A-G-P	7

Figure (2) Applying our proposed algorithm on the news data base and generating frequent item sets



Pass	Freq Itemsets	Supp.
1	A	267
1	G	208
1	P	804
1	T	34
2	A-G	12
2	A-P	26
2	G-P	22
3	A-G-P	7

Figure (3) The generated frequent item sets



min. conf. 80% DBI = 1256

Generated Normal Rules Generate Interesting Rules

Rule	With Confidence
1. G-T → A	100.00 %
2. A-T → G	100.00 %
3. A-T → P	100.00 %
4. G-T → P	100.00 %
5. G-P-T → A	100.00 %
6. G-T → A-P	100.00 %
7. A-P-T → G	100.00 %
8. A-T → G-P	100.00 %
9. A-G-T → P	100.00 %
10. G-T → A-P	100.00 %
11. A-T → G-P	100.00 %

Figure (4) Total interesting and non interesting Rules discovered from the algorithm

The pervious figures show the data set after conversion into the required format of the Proposed Incremental algorithm (PIA) (Figure 1).

The next figure (Figure 2) shows the applying of the algorithm on the converted data base then the generated itemsets according to the specified support and confidence (Figure 3).

The last figure (Figure 4) shows the association rules extracted from the data base with the related confidence.

V. CONCLUSIONS

The proposed incremental algorithm (PIA) applied on the news database has discovered many conclusions. The rules discovered were then made available to the journalists for comment. These rules were very satisfactory and have been approved from their point of view by domain specialists.

A. Conclusion 1 (analysis for the association between topics):

From the previous rules we can discover that for the website to increase its productivity it should be dynamic in the way that according to the user visited links the website should display the topics that are most likely to be visited by the users, this dynamicity can be illustrated by the following table:

If these links are visited	These topics should be displayed as it is most likely to be visited by the users
General and Technical	Athletics and Politics
Athletics and Technical	General and Politics
General and Politics and Technical	Athletics
Athletics and Politics and Technical	General
Athletics and General and Technical	Politics

B. Conclusion 2 (analysis for the most important topic categories):

The most important topic categories that are visited by the users are as the same order as the following table

Politics	1859	67.87%
General	422	15.41%
Athletics	407	14.86%
Technical	51	1.86%

When these results were illustrated on the decision makers in the news paper they decided to increase the politics news to attract more readers.

C. Conclusion 3 (analysis for the most important topics):

The proposed algorithm was used to analyze what IP Addresses vote in which topic. Mining the web site revealed that the order of the most visited categories were political, general, athletic and technical. Furthermore, the most frequent issues of concern to the web site visitors, at most of the topics in all categories, were related to every day cost of living including food and transportation. The news database was converted into the required association rules form generating frequent item sets. Results revealed that in order to increase the number of visitors, the site should be

dynamically changing according to the user interests which were specified when analyzing the association between the topics.

REFERENCES

- [1] Douglas Shona, Agarwal Deepak, Alonso Tirso, Bell Robert, Rahim Mazin, F. Swayne Deborah, Volinsky Chris, "Mining Customer Care Dialogs for Daily News," IEEE Transactions on Speech and Audio Processing, Special Issue on Data Mining of Speech, Audio and Dialog, pp. 652- 660, 200.
- [2] Fatudimu I.T, Musa A.G, Ayo C.K and Sofoluwe A. B, "Knowledge Discovery in Online Repositories: A Text Mining Approach," European Journal of Scientific Research, ISSN 1450-216X Vol.22 No.2, EuroJournals Publishing, Inc., pp.241-250, 2008.
- [3] Feldman Ronen, "Mining Text Using Keyword Distributions," Journal of Intelligent Information Systems 10, Kluwer Academic Publishers. Manufactured in The Netherlands, pp. 281–300, 1998.
- [4] Hung-yu Kao, Shian-hua Lin, Jan-ming Ho, Ming-syan Chen, "Mining web informative structures and contents based on entropy analysis," IEEE Transactions on Knowledge and Data Engineering, VOL. 16, NO. 1, January, 2004
- [5] Laila M. ElFangary, Walid A. Atteya, "Mining Databases by Means of an Incremental Association Rule Learner," International Conference on Convergence and hybrid Information Technology (ICCIT08), IEEE publisher, 2008.
- [6] IAN H. WITTEN, "Adaptive Text Mining: Inferring Structure from Sequences," Journal of Discrete Algorithms, 2(2), pp. 137-159. Elsevier B.V., 2004
- [7] Kjetil Norvag and Randi Oyri, "News item extraction for text mining in web newspapers," Proceedings of International Workshop on Challenges in Web Information Retrieval and Integration in conjunction with ICDE, 2005.
- [8] Louise A. Francis, "Taming Text: An Introduction to Text Mining," Casualty Actuarial Society Forum, Winter 2006.
- [9] Manuel Montes y Gómez, Alexander Gelbukh, and Aurelio López-López. "Mining the news: trends, associations, and deviations," Computación y Sistemas, Vol 5, No. 1, July-September 2001.

- [10] Raymond J. Mooney and Un Yong Nahm, "Text Mining with Information Extraction" , Multilingualism and Electronic Language Management: Proceedings of the 4th International MIDP Colloquium, September 2003, Bloemfontein, South Africa, Daelemans, W., du Plessis, T., Snyman, C. and Teck, L. (Eds.) pp.141-160, Van Schaik Pub., South Africa, 2005.
- [11] Seokkyung Chung, J. Jun, and D. McLeod. Incremental mining from news streams. Encyclopedia of Data Mining and Warehousing. Idea Group Inc., 2005