International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

An Efficient Approaches to Weighted Recursive Pattern Mining Using Association Rules

T. Vijayakumar[1], S.Omprakash[2], T. Senthamarai[3]

[1]Research Scholar, Department of Computer Science, Kovai Kalaimagal College of Arts & Science, Coimbatore-109, India.

[2] Assistant Professor, Department of Computer Applications, Kovai Kalaimagal College of Arts & Science, Coimbatore-109, India

[3] Assistant Professor, Department of Computer Science, St.Paul's College of Arts and Science for Women, Coimbatore-25, India

Abstract: ***Researchers have proposed frequent pattern mining algorithms that are more efficient than previous algorithms and generate fewer but more important patterns. Many techniques such as depth first/breadth first search, use of tree/other data structures, top down/bottom up traversal and vertical/horizontal formats for frequent pattern mining have been developed. Most frequent pattern mining algorithms use a support measure to prune the combinatorial search space. Alternative measures for mining frequent patterns have been suggested to address these issues. One of the main limitations of the traditional approach for mining frequent patterns is that all items are treated uniformly when, in reality, items have different importance. For this reason, weighted frequent pattern mining algorithms have been suggested that give different weights to items according to their significance. The main focus in weighted frequent pattern mining concerns satisfying the downward closure property. Our main approach is to push weight constraints into the pattern growth algorithm while maintaining the downward closure property. We develop WFIM (Weighted Frequent Itemset Mining with a weight range and a minimum weight), WLPMiner (Weighted frequent Pattern Mining with length decreasing constraints), WIP (Weighted Interesting Pattern mining with a strong weight and/or support affinity), WSpan (Weighted Sequential pattern mining with a weight range and a minimum weight) and WIS (Weighted Interesting Sequential pattern mining with a similar level of support and/or weight affinity) The extensive performance analysis shows that suggested approaches are efficient and scalable in weighted frequent pattern mining.***

Keywords: Frequent Pattern Mining, WFIM, WLPMiner, WIP, WIS.

1. Introduction

The World Wide Web (WWW) is fast becoming a rich source for information on people's tastes, dispositions, interactions [1]. The last two generations of humanity have made the Internet an integral part of their everyday lives; from blogging to online shopping to product reviews to social networks. The intent of this thesis is to provide an overview of how the Web can be used for competitive intelligence. In particular, information gathering, information analysis, information verification, and information security.

1.1 Objectives of Data Mining

The objective of the system of developing software for maintaining the activities removes the drawbacks of the earlier system. Computerization Advantages,

- It is easy to retrieve the details more easily and quickly using competitive web mining.
- It is also possible to search the details of the particular domain till date.
- Identifying and summarizing the competitive evidence that details the competitors' strength.
- It is used for easy monitoring facility.

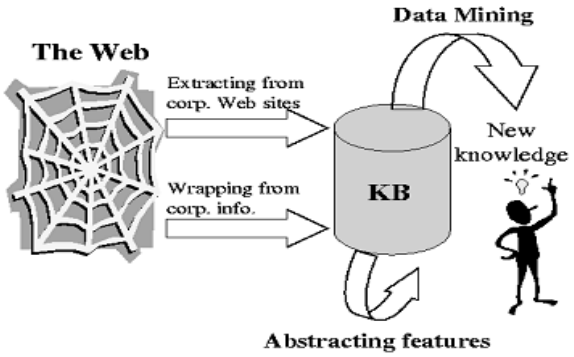

Figure 1.1: Data Mining Process

This thesis is concerned with the problem of mining competitors from the Web automatically. The task of competitor mining that we address in the project includes mining all the information such as competitors, completing domains, and competitors' strength. A novel algorithm called CoMiner is proposed, which tries to conduct a Web-scale mining in a domain independent manner.

The CoMiner algorithm consists of three parts:

- Given an input entity, extracting a set of comparative candidates and then ranking them according to comparability.
- Extracting the domains in which the given entity and its competitors play against each other.
- Identifying and summarizing the competitive evidence that details the CoMiner algorithm is presented. The

experimental results show that the proposed algorithm is highly effective

2. Review of Literature

The rapid development and its relative maturity, the modern Web becomes a sensor of the real world and records the real world from many aspects everyday [3]. Wrapper-based approaches have also been proposed for extracting information from highly structured documents.

There are also some work about the comparative search and mining. Liu'sworkis [7] mining opinions and extracting sentiment from some online discussion forms. Zhai [9] defines a comparative text mining problem(CTM) which means discovering common themes and specific theme for an existing set of comparative text collections. The set of comparative data automatically from the Web. Sun's [8] comparative search engine, collects comparative information for the given two entities.

Many methods for entity recognition and extraction have been proposed, most of which are focused on the use of supervised learning techniques such as Hidden Markov Models [4]. Wrapper-based approaches have also been proposed for extracting information from highly structured documents [5].

2.1 Data Mining Review

Data Mining: Today, size of databases can be very large. Within this data you can find hidden strategic information. But when you have a huge amount of data, inducing meaningful conclusions is not easy. The novel answer is data mining being used both to increase revenues and to reduce costs. Many people use data mining as a synonym for another popular word, Knowledge Discovery in Database KDD.

The KDD processes are shown in Figure 2.1 (Han, J., & Kamber, M.,) [2]. The preprocessing includes data cleaning, integration, selection and transformation.

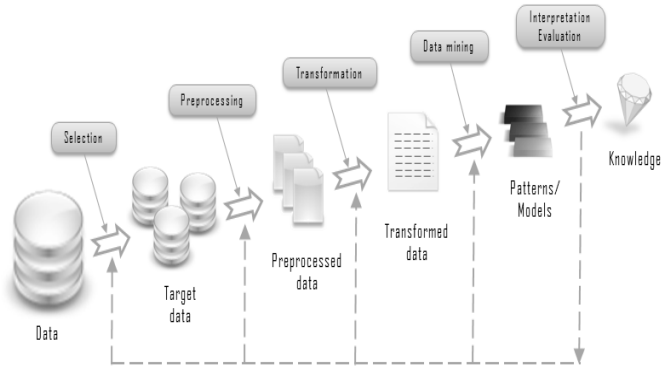

Figure 2.1: Knowledge Discovery in Database Processes

(Song et al) [6] defines data mining as a process of exploration and analysis of large quantity of data to discover meaningful patterns and rules.

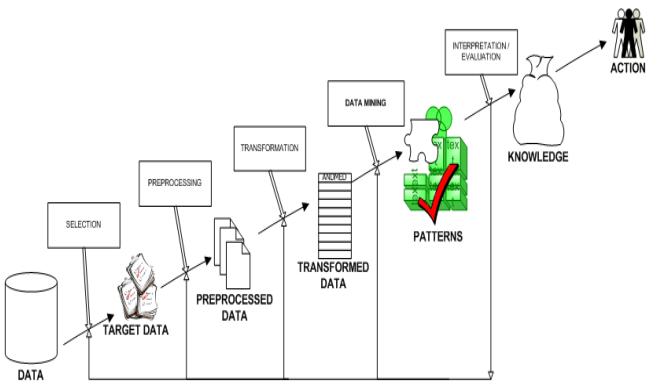

Figure 2.2: The Major Steps in Data Mining Process

Functions of Data Mining: (Dunham) categorizes data mining to two categories, one is descriptive and the other one is predictive (Figure 2.3).

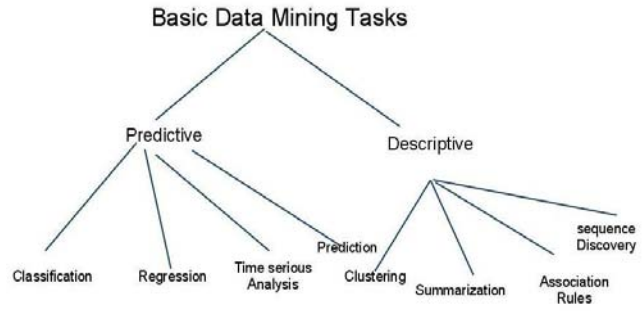

Figure 2.3: Classification of DM Techniques

The most important step is right data selection, data gathering and data exploration. It must build a predictive model based on patterns determined from known results, and then examine that model with a new sample data. Nearest neighbor methods are also talked about in many statistical texts on classification, such as (Duda and Hart cited by Han, J., & Kamber, M) [2] and (James cited by Han, J., & Kamber, M) [2].

Clustering: Classification can be taken as supervised learning process, clustering is another mining technique similar to classification. "Clustering is the process of grouping a set of physical or abstract objects into classes of similar objects" (Han, J., & Kamber, M)[2].

2.2 Association Rules

Association rule mining is one of the most important techniques of data mining. (Agrawal et al) [10] Introduced this method first time. The goal of this technique is extracting interesting correlations, frequent patterns, and associations among sets of items in the transaction databases.

Web Usage Mining: Web usage mining is the application that uses data mining to analyze and discover interesting patterns of user's usage data on the web.

Web Content Mining: Web content mining is the process to discover useful information from text, image, audio or

video data in the web. Web content mining sometimes is called web text mining, because the text content is the most widely researched area.

Web Structure Mining: Web structure mining is the process of using graph theory to analyze the node and connection structure of a web site. According to the type of web structural data, web structure mining can be divided into two kinds.

Web Mining: Web mining is the application of data mining techniques to discover patterns from the Web. According to analysis targets, web mining can be divided into three different types, which are Web usage mining, Web content mining and Web structure mining.

2.3 Pros of Web Mining

Web mining essentially has many advantages which makes this technology attractive to corporations including the government agencies. The government agencies are using this technology to classify threats and fight against terrorism. The predicting capability of the mining application can benefits the society by identifying criminal activities.

2.4 Cons of Web Mining

The obtained data will be analyzed, and clustered to form profiles; the data will be made anonymous before clustering so that no individual can be linked directly to a profile.
The growing trend of selling personal data as a commodity encourages website owners to trade personal data obtained from their site.

2.6 Content Analysis

Ole Holsti groups 15 uses of content analysis into three basic categories:

- Make inferences about the antecedents of a communication
- Describe and make inferences about characteristics of a communication
- Make inferences about the effects of a communication.

3. Research Methodology

The task of competitor mining that we address in the paper includes mining all the information such as competitors, competing domains, and competitors' strength. A novel algorithm called CoMiner is proposed, which tries to conduct a Web-scale mining in a domain independent manner. The CoMiner algorithm consists of three parts:

- Given an input entity, extracting a set of comparative candidates and then ranking them according to comparability
- Extracting the domains in which the given entity and its competitors play against each other
- Identifying and summarizing the competitive evidence that details the competitors' strength.

3.1 Candidate Extraction

We define a set of linguistic patterns for getting the pages which may contain information of competitors and for extracting candidate's competitor. The last two patterns are often used to compare two entities. En refers to Entity Name and CN refers to Competitor Name.

H1: such as EN (, CN)*or and CN

e.g., "brands of tape such as Sony, Phillips, BASF or TDK should be used."

3.2 A Text Mining Strategy for Competitive Intelligence

The concept-based approach allows qualitative and quantitative analyses on the content of a textual collection. The strategy is segmented in the following steps:

Pre-processing (text retrieval and data normalization);

- Concept extraction;
- Pattern mining;
- Definition and execution of rules to extract relevant data for each concept;
- Evaluation and analysis of the results for CI.

Pre-processing: This step combines data retrieval and data cleansing sub-processes. The first is used to collect (retrieve) the texts according to their source and type.

3.3 Pattern Mining

The goal of this step is to find interesting patterns in concepts distributions inside a collection or sub-collection. A used technique is the concept distribution listing, which analyses concept distributions in a group of texts (in the whole collection or in a sub-collection).

A software tool counts the number of texts where each concept is present, generating a vector (called centroid) of concepts and their frequencies (or proportion) inside the group.

3.4 Research Design

A research design is a roadmap for performing the marketing research project. It gives details of each step in the marketing research project. Accomplishment of the research design should result in all the information requested to construction or solve the management-decision problem (Malhotra, K.N)[11].

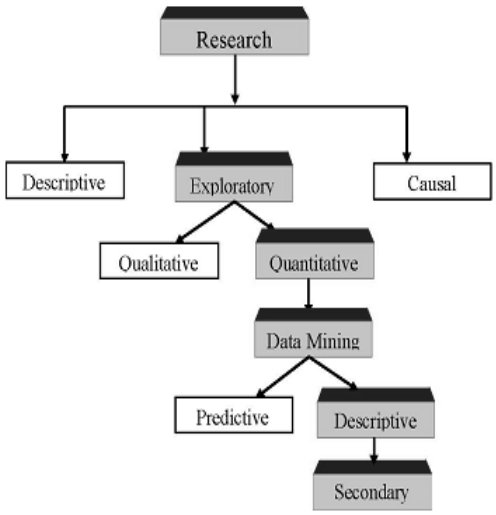

Figure 3.1: Research Design of this Study

3.5 Research Process

The purpose of this research is to understand changes happening in the customer buying behavior during time. Figure 3.2 shows the general overview of change mining flowchart.

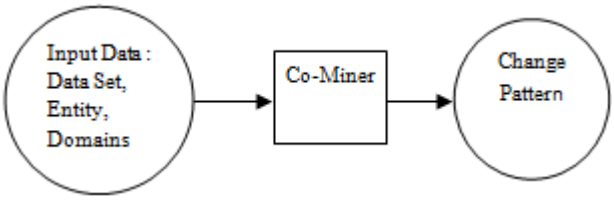

Figure 3.2: Change Mining Process Perspective

4. Experimental Results

Analysis of Evidence Mining with the Web Recall that competitive evidence is defined as a sentence that contains competitive information. Typical competitive evidence contains four elements:

- Entity EN queried by the user,
- Entity CN automatically discovered by CoMiner at step1,
- Competitive domain D specified by the user or automatically identified by CoMiner at step2, and
- The competitive relation indicating the comparative type between EN and CN.

4.1 Modules

a) Identifying the List of Competitors & Competitive Domains
b) Extraction of Entity & Key Phrase
c) Competitor Ranking
   - Computation of Match Count
   - Computation of Mutual Information
   - Computation of Candidate Confidence (CC) & Modeling of Ranking
d) Filtering of Synonyms & Domain Names
e) Competitive Evidence Mining

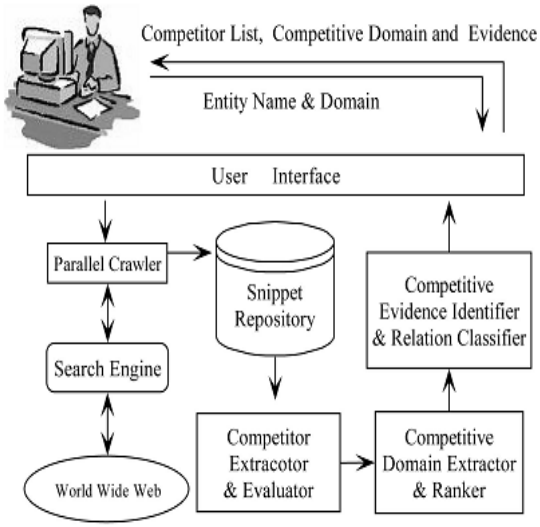

Figure 4.1: CoMiner Prototype System Architecture

4.2 Identifying the List of Competitors and Competitive Domains

The objective of this step is to extract and then to rank the competitors of the given entity from a set of pages returned by the search engine.

Web Redundancy: Although on the Web there are lots of varied expressions that indicate the comparative relationships between the given entity and its competitors, we need only a few common patterns to extract candidates from the web pages.

4.3. Extraction of Entity & Key Phrase

To extract the competitive domains for different pairs of competitors, we need to acquire a better understanding of the complicated distribution of competitive domains since varied domains rarely share common patterns.

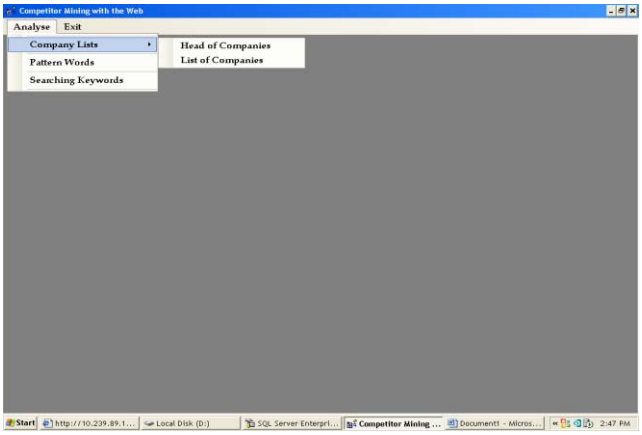

Figure 4.2: Extraction of Competitor Mining

4.4 Competitor Ranking

It is first send the query containing both the query entity and a competitor name to the search engine. Then, we use the top 100 returned results as our data set for extracting

competitive domains. E.g., phrase frequency (PF), document frequency (DF), and average distance (AD)).

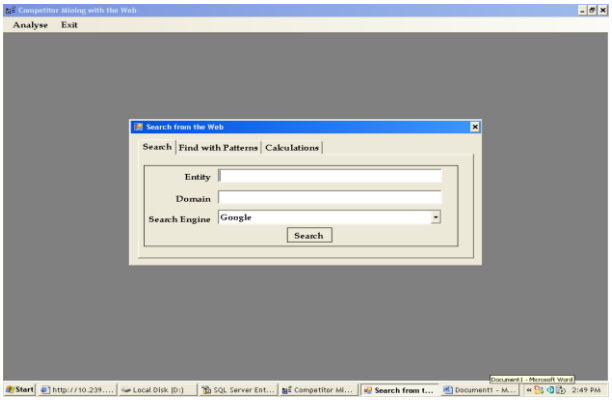

Figure 4.3: Competitor Mining with Web

The features are described as follows:

- **PF:** This feature is calculated in the traditional meaning of term frequency (TF). In general, a frequent phrase is more likely to be a good candidate of salient phrase.
- **DF:** This feature is measured by the number of documents containing the phrase. If a phrase appears in most pages containing both the entity name and its competitor, it may have a high probability to be a competitive domain.
- **PL:** Intuitively, a longer name is more meaningful for user's browsing.
- **AD:** This feature is calculated by the distance between the phrase and the given entity or its competitors.

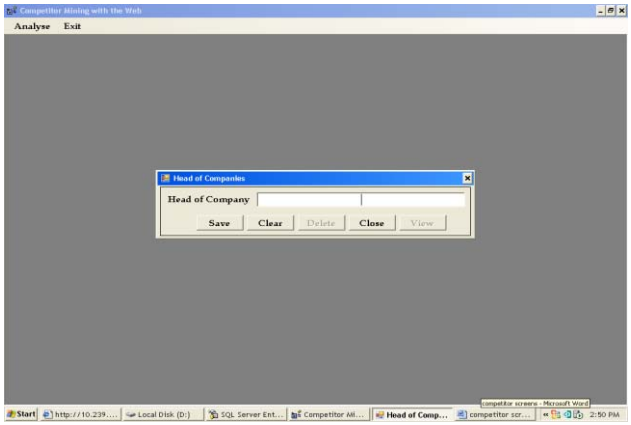

Figure 4.4: Head of Competitors

Match Count: A candidate's match count (MC) is calculated as the number of times it is extracted from the result set by our predefined patterns. Intuitively, the more times the candidate is matched, the more comparative the relationship between the candidate and the given entity is:

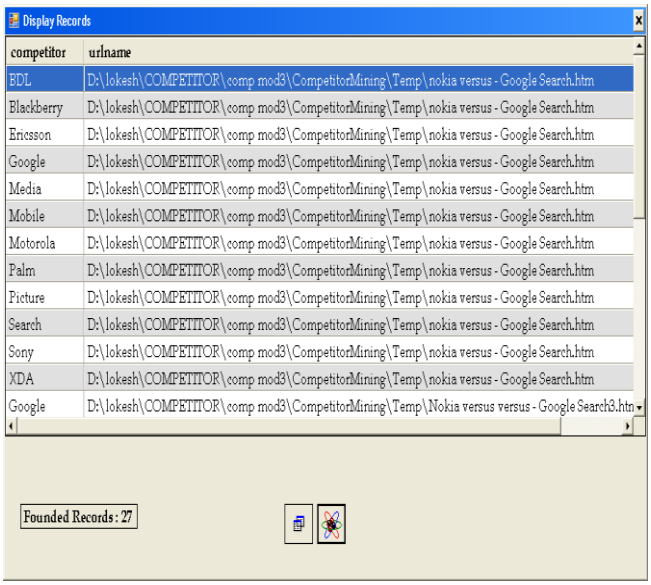

Figure 4.5: Displaying Records

- **MC(c,e) = Summation(p belong P) Count(c,e,p),** where MC(c,e) means the hits of all extraction patterns; another formula for calculating this feature is to linearly weight the contribution of each pattern for calculating MC(c,e) :
- **MC(c,e) = Summation(p belong P) w(p) Count (c,e,p)** where wp is the weight of pattern p. Intuitively, we give patterns C1 and C2 higher weights since they express more competitive meanings. In our experiment, the weights of C1 and C2 are set to five, while others are all set to one.

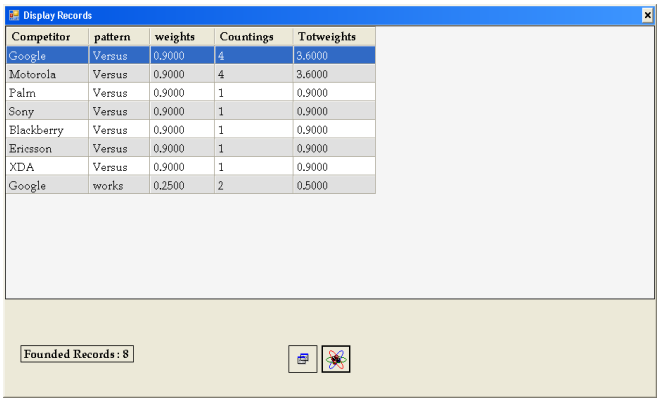

Figure 4.6: Match Count

Mutual Information: The point wise mutual information (PMI) is often used to measure the co-occurrence between two terms. Here, we use it as a feature to measure the comparability between the given entity and its competitor since the more frequently they co-occur, the more related they are to each other:

$$\mathbf{PMI\ (c,e) = hits(c,e) / hits(c)\ hits(e)}$$

where hits represents the number of returned search results that contain x. where hits(x) represents the number of returned search results that contain x.

4.5 Competitive Evidence Mining

To have a better understanding of competitors, in this step, we mine the detailed competitive evidences from the gathered descriptions it is recalled that competitive evidence is defined as a sentence that contains competitive information. Atypical competitive evidence contains four elements:

- Entity EN queried by the user,
- Entity CN automatically discovered by CoMiner at step 1,
- Competitive domain D specified by the user or automatically identified by CoMiner at step 2, and
- The competitive relation indicating the comparative type between EN and CN with the help of Web redundancy,

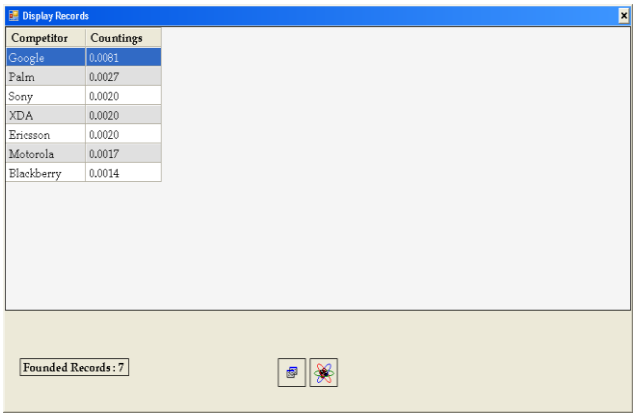

Figure 4.7: Competitive Evidence Mining

System Setup : To evaluate the effectiveness of CoMiner, a prototype System is implemented. During the implementation, the system's response speed becomes our major consideration.

Tables 4.1: Training Data Distribution

Field	Num	Training Entity
Company	(2)	Morgan Stanley, Citigroup
University	(2)	Cambridge, Oxford
Basket Ball	(2)	Michael Jordan, Shaquille O'Neal
Messenger	(2)	Skype, ICQ
Watch Brand	(2)	Rolex, Citizen

Data Preparation: We first manually collected 10 entity queries. The 10 entity queries are from five categories: Company, University, Basketball, Messenger, and Mobile Cell Phones. We select famous IT companies, football stars, and popular products so that the results can also be evaluated easily by more readers.

Competitor Discovery: We further manually label the top30 returned pages for each of the 70 entities to check out whether the returned pages contain the competitor names.

Table 4.2: Test Data Distribution

Field	Num	Input Entity Example
IT Company	(10)	Microsoft, Google, Sony,
Cell Phone	(5)	Nokia, Motorola, Siemens
Digital Camera	(5)	Cannon, Nikon, OLYMPUS
Computer	(10)	Toshiba, DELL, IBM
Brand of Car	(10)	BMW, Benz, Audi
Product	(5)	Motorola V360, Canon A70,
Football Star	(5)	Ronaldo,Zidane,Lampard
University	(10)	Princeton, Yale,Cornell,
Football Club	(10)	AC Milan, Arsenal,Liverpool

5. Conclusion

Mining competitive information has attracted a great amount of attentions in recent years. The main contributions are the following:

- The observation of competitor, competitive domain, and competitive evidence distribution in the unrestricted WWW,
- The proposal of a novel algorithm, CoMiner, which can effectively mine competitor information from the Web, and
- The implementation of the CoMiner and the experimental results showing that the proposed algorithm is highly effective.

6. Scope for Future Work

In our future work we plan to evaluate CoMiner in more domains and improve the CoMiner to mine more competitive information from the Web. In addition, it is quite interesting to find out that Kaifu-Lee was extracted as a competitive domain between them, perhaps owing to the lawsuit case for him that the two giants are involved in. Besides the traditional rivals, Oracle in database and Adobe in office, etc., we find some unexpected information. For example, spam is considered as a rival because of its adverse influence on the customer's Using Hotmail.

References

[1] T.L.Friedman, The World Is Flat: A Brief History of the Twenty-First Century. Farrar, StrausandGiroux,Apr.2005.
[2] Chen, M.C, Chiub, A.L, Chang, H.H, (2005), Mining changes in customer behavior in retail marketing, Expert System with Applications, Volume 28, Issue 4, pp.773781
[3] Q.Zhao, T.-Y.Liu, S.S.Bhowmick, and W.-Y.Ma, "Event Detection from Evolution of Click-Through Data," Proc.ACM SIGKDD'06, pp.484-493, 2006.
[4] D.Freitagand A.McCallum, "Information Extraction with HMMS and Shrinkage," Proc.AAA I Workshop Machine Learning for Information Extraction, 1999.
[5] N.Ashish and C.Knoblock, "Wrapper Generation for Semi Structured Internet Sources," SIGMOD Record, pp.8-15, 1996.
[6] Taoying Li and Yan Chen, "Fuzzy ClusteringEnsemble with Selection of Number of Clusters", Journal of Computers, Vol. 5, No. 7, Pp. 112 119, 2010.

[7] M.Hearst, "Automatic Acquisition of Hyponyms from Large Text Corpora," Proc.14thInt'lConf.Computationa Linguistics (COLING'92), pp.539-545,1992.
[8] P.Cimino S.Handschuh, and S.Staab, "Towards the Self Annotating Web," Proc.13th Int' Conf.World Wide Web (WWW'04), pp.462-471,2004.
[9] B.Liuand C.Chin, "Mining Topic-Specific Concepts and Definitions on the Web,"
[10] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy, "Advances in Knowledge Discovery and Data Mining", MIT Press, Cambridge, Mass., 1996.
[11] V. Chandola, S. Boriah and V. Kumar, "Similarity measures for categorical data -A comparative study", Technical Report 07-022, Department of Computer Science & Engineering, University of Minnesota, 2007.

Author Profile

Mr. T. Vijayakumar., M.Sc., M. C. A., M. Phil., Research Scholar, Department of Computer Science

Mr. S. Omprakash., M.Sc., M. Phil., Assistant Professor, Department of Computer Science

Ms. T. Senthamarai, M. Sc., M. Phil., Assistant Professor, Department of Computer Science