IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

Purchasing Behavior of the Customer Using Co-Miner Algorithm

S.S.Suganya[*1], R.TamilSelvi[2], K.Sathyapriya[3],R.Suganya[4]

[*1,2,3,4] Department of Computer Science, Dr. SNS. Rajalakshmi College of Arts & Science, Coimbatore-49, TamilNadu, India

Suganya.annur@gmail.com

Abstract

In the extended world, the competition is becoming more and more sweltering in various domains comparing with similar products that have the same features to the consumers who are about to purchase a certain product. The paper consist tasks involving the process of mining Competitive entities, domains, evidences for decision making to be competitor mining problems. The task of competitor mining that address in the paper includes mining all the information such as competitors, competing domains, and competitors' strength. An algorithm CoMiner is suggested, which tries to conduct a Web-scale mining in a domain in dependent manner. The algorithm contains:

1) Given an input entity, extracting a set of comparative candidates and ranking them.
2) Extracting the domains in which the given entity and its competitors play against each other.
3) Determining and summarizing the competitive evidence that details the competitors' strength.

The competitor mining tasks includes acquiring the competitors for a given entity, elaborating the competitive domains with respect to the competitors, and summarizing the opinion of detailed competitive evidences. Although on the Web there are many separate expressions that denote the comparative relationships between the given entity and its Competitors, we need a few patterns to extract candidates from web pages.

Keywords: Extraction, Web-scale mining, competitive evidence.

Introduction

The World Wide Web (WWW) is fast becoming a rich source for information on people's tastes, dispositions, interactions. 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.

As this information is in the public domain, it represents a rich source of context on people, their interactions and their habits. The opportunity to leverage this is very tempting for most corporations. For others, this paradigm shift represents an indication of where their business should be going.

The intent of this thesis is to provide an overview of how the Web can be used for competitive intelligence. Following a definition of Competitive Intelligence, the logical structure of the World Wide Web is reviewed to provide a foundation for understanding how information is stored on and retrieved from the web and the difficulties that arise from using this logical approach. Sections that follow detail the techniques that can be used to carry out CI projects and some of the problems associated these techniques. In particular, information gathering, information analysis, information verification, and information security

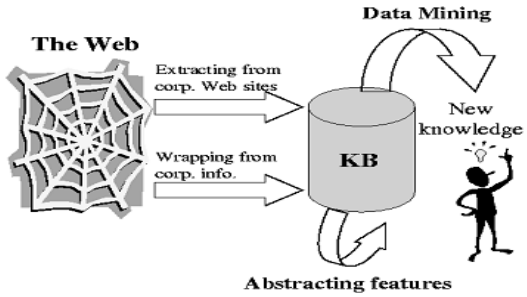

Figure 1.1 Data mining Process

Objective

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

Computerization Advantages,

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

Salient Features of the System

This thesis is concerned with the problem of mining competitors from the Web automatically. Nowadays, the fierce competition in the market necessitates every company to know not only which companies are its primary competitors but also in which domains the company's rivals compete with itself and what its competitor's strength is in a specific competitive domain. The task of competitor mining that we address in the project 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:

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

Literature Study

This chapter provides brief details about the background study that was successfully carried out as a part of the system study. This section discusses about the existing and their short comings.

With 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 every day. As a result, it contains all kinds of comparable information and provides an ideal platform to conduct competitor mining. However, we are confronted with the abundance of the Web-scale information distributed in various forms (e.g., bbs, blog, and mailing list), which means too many alternatives. One method to deal with the competitor mining problem is based on the use of search engines

For a given entity, one can submit a query representing the entity, and the search engine will return a set of related web pages. The competitor information might be obtained via browsing through the related pages. However, such a browsing process is obviously tedious because only a few pages may contain the comparative information among thousands to browse, especially in the case that one knows little about the entity. Furthermore, it is also hard for the method to specify the competing field and to summarize the opinions of all identified competitive evidences. There are also some services available to help people get the competitive information of a given product. Websites like Froogle3 and Amazon4 have already provided product comparison services such as publishing the prices and customer reviews. These sites, however, are designed to serve for a limited domain, and their services are based on manually built databases.

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. These methods require a large set of training data, and the results are usually affected by the similarity between the test data and the training data. Wrapper-based approaches have also been proposed for extracting information from highly structured documents. Hardly is it, however, suitable for general web pages, because they lack of a common structure. Meanwhile these methods are usually proposed for recognizing entity names in a web page rather than finding particular entity names related to a given entity.There are also some work about the comparative search and mining. The mining opinions and extracting sentiment from some online discussion forms. It defines a comparative text mining problem(CTM) which means discovering common themes and specific theme for an existing set of comparative text collections .But our work uses a web search engine to get a set of comparative data automatically. Comparative search engine, collects comparative information for the given two entities. Our work, however, automatically discover highly competitive entities instead.

Entity Extraction

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. These methods require a large set of training data, and the results are usually affected by the similarity between the test data and the training data. Wrapper-based approaches have also been proposed for extracting information from highly structured documents .Hardly is it, however, suitable for general web pages, because they lack of a common structure. Mean while these methods are usually proposed for recognizing entity names in a web page rather than finding particular entity names related to a given entity.Some pattern-based approaches have also been proposed. Quite a few different kinds of problems can be solved via such pattern-based approaches.

Key Phrase Extraction

Our work on extracting competitive domain from web pages is related to previous studies of identifying key (salient)phrases in text mining .Many features have been proposed. One popular used property is Term Frequency/ Inverted Document Frequency (TFIDF) The independence of a phrase is

also proposed and measured by the entropy of its context More properties like phrase length (PL).

Comparative Search and Mining

Much work has been done on helping companies and individuals gain marketing information by mining online resources.

Research Methodology

System Proposed

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. 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. In other words, the entities distributed in the infinite domains can be extracted with the use of finite kinds of patterns. To extract the competitor names for a given entity, we define a set of linguistic patterns for acquiring the pages that may contain information about competitors. For extracting competitors, we sent a series of queries constructed from the entity name and predefined patterns. We propose to compute Match Count, Mutual Information, and Candidate Confidence (CC). Based on these values we calculate Confidence Score (CS). 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. 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. After collecting the pages, we use an NLP tool to parse the results and get a list of noun phrases as the candidate list for extracting domain names. The experimental results show that the proposed algorithm is highly effective.

The objective of this step is to extract and rank the competitors of the given entity from a set of pages. Our competitor discovery algorithm is based on the following observation.

Observation

The expressions which indicate the comparative relationship are diversified, but we need only few common patterns to extract candidates due to the web redundancy. Un even Co-occurrence also gives the measurement of the degree of the relationship closeness. Comparative entities often occur together more frequently.

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 first 3 kinds of patterns have been used by Hearst to identify is-a relationship between the concepts referred by two terms. However, the two terms are usually the competitors. The last 2 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."

A Text Mining Strategy for Competitive Intelligence

The main decision was to use an approach accessible to small and medium-sized enterprises in developing countries, without the need to buy expensive information retrieval software. The concept-based approach allows qualitative and quantitative analyses on the content of a textual collection. Qualitative analysis identifies concepts present in the texts and quantitative analysis extracts patterns in concepts distributions through statistical techniques. Then, comparisons help to identify different characteristics that can be used as competitive advantages. 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. If the quantity of documents is large it is possible to select only a sample of them, using statistical Sampling techniques. After the documents are collected it is necessary to clean them to exclude unnecessary information like prepositions, articles, conjunctions, adverbs and other frequently used

words, plus ETO domain specific and structural words.

These words are known as stop words and they are all removed from the texts.

Concept Extraction

The purpose of this step is to identify concepts present in texts. However, documents do not have concepts explicitly stated, but instead they are composed of words that represent the concepts. As concepts are expressed by language structures, it is possible to identify concepts in texts analyzing phrases.

Considering that many concepts can be found in a document and that the user has specific needs, the first thing to do is to define the concepts that are relevant to the user. He must express which words and expressions indicate the presence or absence of each concept he is interested to extract from the ETO System. To help him in this process, it is necessary to analyze some ETO that belong to the context of the user's business and look for the most frequent words of a document or a set of documents. The idea is that documents belonging to the same cluster have the maximum probability of having the same concept(s). Thus, their words have also a high probability of belonging to the same concept(s). The concept of a cluster can be identified by analyzing the most frequent words in the cluster. The other words in the centroid can be used as concept descriptors. After the concepts are identified they can be expressed and modeled by rules. Rules combine positive and negative words that indicate the presence or the absence of concepts in a phrase. The goal is to verify whether a concept is mentioned in a phrase. If the concept is present more than once in a text, the total counting is used to define an associative degree between the text and the concept, indicating how much a concept is referred by a text.

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. This technique allows finding what dominant themes exist in a group of texts. Also we can compare one centroid to another (between different groups) to find common concepts in different groups or to find variations in distributions of a certain concept from one group to other. Another possible usage is to find differences between groups, that is, concepts present in only one group (exclusive concepts).

Another used technique is the association or correlation. It discovers associations between concepts and expresses these findings as rules in the format X ! Y (X may be a set of concepts or a unique one, and Y is a unique concept). The rule means, "if X is present in a text, then Y is present with a certain confidence and a certain support". Following the definitions of Lin et al. and Garofalakis et al., confidence is the proportion of texts that have both X and Y compared to the number of texts that have only X, and support is the proportion of texts that have both X and Y compared to all texts in the collection. Confidence is similar to the conditional probability (if X is present, so there is a certain probability of Y being present too).

This allows predicting the presence of a concept according to the presence of another one.

Rules to Extract Relevant Data Associated With Concepts

Each message belonging to the concept selected by the user may contain much relevant information,. Each concept may have specific information to be extracted. This necessity of information is specific to this particular concept. When this concept is identified in a message this information should be extracted. The user may have different needs of information for different concepts.

The creation of rules for extracting different types of information for each concept is the central point to the success of the extraction process. The rules must be defined and constructed by the user with the aid of an expert in the field of his business, After the identification of association of words is performed, the user must examine the results looking for words or concepts associated to the concept The idea behind this method is that associations that have high support (frequently occur) indicate that almost all documents within the concept chosen have the associated information and this information may be important. Of course this is only a technique used to aid the rules construction process. The expert must do the final decision and refinement.

It is important to state clearly that these rules, beyond retrieving the most important and relevant information to the user, according to the context of the message (its concepts), allow the user to read only the subset of messages considered to be related to the concepts of interest. This is extremely interesting in case the number of messages is very high. In this case, only the most important information of each message is extracted and the user obtains a summary or a report of the information received.

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)[34]. Many designs maybe are suitable for a given marketing research problem. A good research design ensures that the information gathered will be related and useful to management and that all of the necessary information will be achieved. A good design should also assist to ensure that the marketing research project will be performed effectively and efficiently (Malhotra, K.N)[34]. The research design of this study is illustrated in

figure 3.1. Detailed descriptions are explained below.

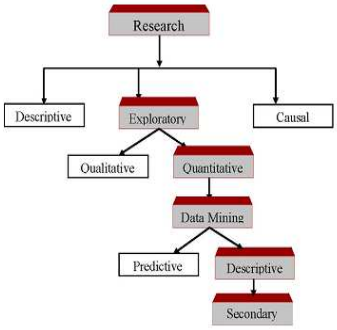

Figure 3.1 Research design of this study

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.

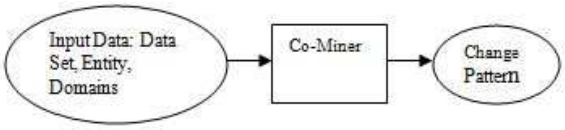

Figure 3.2 Change mining process perspectives

As it is shown, the input of this flowchart is data which show the customer purchasing behavior, some demographic variables and product data. This data induced to the Change Miner. Change mining procedure consists of different steps implemented by different data mining techniques and algorithms in each. In this chapter, different studies related to change mining were reviewed. Based on literature, change mining has several steps, includes describing customer behavior by mining association rule and mining change pattern.

In this paper, the research process has been followed. The following process is based on previous methodology on Change mining.

The whole process of change mining is shown in figure 3.2. The process consists of several steps such Segmentation, Mining Customer Behavior, steps such as Data Collection, Data Pre-Processing, Customer Mining Customer Behavior, and Change Processing, Customer Mining.

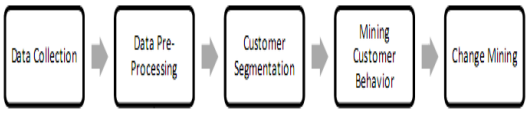

Figure 3.3 Change mining process

Each step by itself consists of several illustrated. Each step by itself consists of several tasks. In figure 3.4 the detail of each step is the detail of each step is

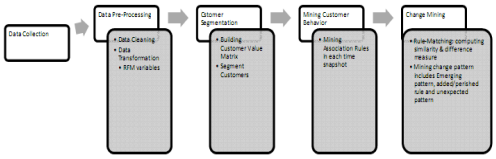

Figure 3.4 Change mining process in detail

Results & Resolutions

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:

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

At the first glance, the mining problem is quite challenging because of

Query diversity. The CoMiner allows the user to input entity queries in various domains, e.g., Ronal do in the domain of football and Canon A70 in digital camera. Different domains have different comparative styles. For instance, 1-1draw is commonly used in the football team comparison. EN is similar to CN are popularly used in product comparisons. Web noise. The snippets gathered from the search engines are dynamically generated and may be composed of incomplete sentences

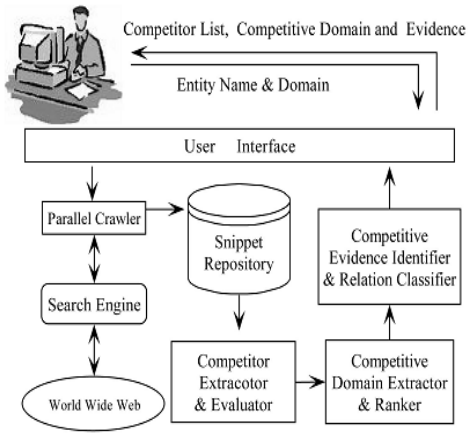

Figure 4.1 Prototype system architecture of CoMiner

Elements

1) Identifying the List of Competitors & Competitive Domains
2) Extraction of Entity & Key Phrase
3) Competitor Ranking
   a. Computation of Match Count
   b. Computation of Mutual Information
   c. Computation of Candidate Confidence (CC) & Modeling of Ranking
4) Filtering of Synonyms & Domain Names
5) Competitive Evidence Mining

Descriptions on Elements

Identifying the List of Competitors & Competitive Domains The objective of this step is to extract and then to rank the competitors of the given ntity 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. In other words, the entities distributed in the infinite domains can be extracted with the use of finite kinds of patterns.

Uneven Co-occurrence: It means that the entity and its competitor usually have much more co-occurrence than the non-competitor pairs

For the input entity, a set of comparative queries are generated by using a set of linguistic patterns

For the comparative queries, a set of informative pages will be gathered by the search engine. The competitor list CL can then be extracted from the informative pages IP.

1 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. The pages that contain both the given entity and its competitor names often talk about their competitive domains. The phrase referring to the competitive domain also is a salient phrase in the data set. By querying the search engine with each pair of names including both the given entity E and its competitor Ci, a set of pages describing their relationship can be obtained. Then, a salient phrase ranking algorithm is used to extract the competitive domains.

2 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. After collecting the pages, we use an NLP tool to parse the results and get a list of noun phrases as the candidate list for extracting domain names. To find meaningful domain names and perform an accurate ranking of the candidates, we improve the existing salient phrase ranking methods in other applications by adding new features. e.g., phrase frequency (PF), document frequency (DF), and average distance (AD). We denote the current phrase as p, and the collection of returned results for the given entity and one of its competitors as C(e,ci) where e represents the given entity, and ci denotes one of its competitors.

The features are described as follows:

1. PF. It 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.
2. DF. It 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.
3. PL. Intuitively, a longer name is more meaningful for user's browsing.
4. AD. This feature is calculated by the distance between the phrase and the given entity or its competitors

2.1 Computation of 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:

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.

2.2 Computation of Mutual Information:

The pointwise 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: 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.

2.3 Computation of Candidate Confidence (CC) & Modeling of Ranking:

There may be some common words in our candidate list, such as company or Web, which may be improperly extracted owing to either the informal expression on web pages or the lack of recognition capability of our defined patterns toward complicated natural language. In all such cases, these words usually appear with a low match count but high frequency in a data set. We use CC to recognize these words. The CC is calculated by the following formula:

$$CC(e/ c) = MC (c,e) /hits(c)$$

where hits (c) represents the frequency of c in the search results returned by the search engine.Given the features above, we use the following formula to combine them and calculate a confidence score (CS) for each competitor:

$$CS(ci) = w1.MC(ci,e) + w2.PMI(ci,e) + w3.CC(e/ ci)$$

The weights are tuned with a Hill-Climbing algorithm based on a set of manually labeled training data.

Filtering of Synonyms & Domain Names

We need to select one from them as our final result to the user. Letting cs denote the short term (e.g., Universal) and cl denote the long term (e.g., Universal Studio), we use the mutual information between cs and cl to decide which term to select:

$$PMI (cs,cl) = hits(cs ,cl) / hits(cs) hits(cl)$$

where hits(c) represents the set of returned results that contains c.If PMI (cs,cl) is larger than a threshold, the long term will be selected.

The objective of this step is to extract the competitive domain for each pair of the given entity and its competitor, e.g., to identify the competitive domain between Microsoft and Sony. 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. Yet by our observation, the extraction can be solved by a salient phrase ranking method based on existing data mining techniques.

Entity-domain co-occurrence: The pages that contain both the given entity and its competitor names often talk about their competitive domains. For example, the pages containing Sony and Microsoft often talk about game console but not digital camera. Yet the pages containing Canon and Sony often focus on digital camera. . Salient phrase is the one that is referring to the competitive domain also is a salient phrase in the data set. For example, game may have more frequency than other terms in pages containing Sony and Microsoft.

Parts of speech (POS) of domain are the meaningful domain names are more likely to be noun phrases.

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. A typical competitive evidence contains four elements:

1. entity EN queried by the user,
2. entity CN automatically discovered by CoMiner at step 1,
3. competitive domain D specified by the user or automatically identified by CoMiner at step 2,4. The competitive relation indicating the comparative type between EN and CN with the help of Web redundancy,

We take a simplified yet effective approach that only utilizes a finite set of general competitive indicators to identify the competitive sentences and competitive relations. The occurrence of entities may help to filter out the irrelevant competitive evidences. The sentences without a target entity can be filtered out directly. With the explored competitive evidences and relations, we can have an overall understanding of the competitive status. Assuming that a set of competitive evidences SE = {E1; E2; . . . ; EM} is mined from M sentences, we define the Positive Comparative Degree (PCD) and Negative Comparative Degree (NCD) as follows:

$$PCD = Mod(S+E) / M, NCD = Mod(S-E) / M$$

where S+E means a subset of SE where evidence indicates that EN is better than CN, and S-E is defined similarly. Both PCD and NCD are normalized into a scale of 0-10 for the ease of understanding

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. The results are below Table. Through the experiment, we can find that our approach increases the number of discovered informative page by 51 percent, and most of the informative pages are found with CP1 and CP2 patterns. Here, "Original" means the way of obtaining informative pages without any patterns. CP represents the Comparison Patterns, and HP represents the Hearst Patterns defined.

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

Table 1 Test Data Distribution

Conclusion

Mining competitive information has attracted a great amount of attentions in recent years. It is not only useful for a company to analyze its rivals in the domains it engaged in but also helpful for a common user to select the right commodity among various choices. In this project, we address a new problem of mining competitors from the general Web automatically. 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.

In our future work, we plan to evaluate CoMiner in more domains and improve the CoMiner to mine more competitive information from the Web.

References

[1] Song et al, H. S., Kim, J. K., & Kim, S. H. (2001). Mining the change of customer behavior in an internet shopping mall. Expert System with Applications.

[2] A.R.Johnson, "What Is Competitive Intelligence?" http:// www.aurorawdc.com / what is ci.htm, 2007.

[3] O.Zamir and O.Etzioni, "Grouper: A Dynamic Clustering Interface to Web Search Results,"

[4] S.Morinaga, K.Yamanishi, K.Tateishi, and T.Fukushinna, "Mining Product Reputations on the Web," Proc.ACM SIGKDD'02, pp.341-349,2002.

[5] M.Hu and B.Liu, "Mining and Summarizing Customer Re-views," Proc.ACMSIGKDD'04, pp.168-177, 2004.

[6] B.Liu, M.Hu, and J.Cheng, "Opinion Observer: Analyzing and Comparing Opinions on the Web,"pp.342-351, 2005.

[7] N.Jindaland B.Liu, "Identifying Comparative Sentences in Text Documents," Proc.ACMSIGIR'06, pp.244-251,2006.

[8] N.Jindal and B.Liu, "Mining Comparative Sentences and Relations," Proc.21st Nat'l Conf. Artificial Intelligence (AAAI), 2006.

[9] B.Liu, Web Data Mining: Exploring Hyper links, Contents and Usage Data. Springer, Dec.2006.

[10] Han, J. & Kamber, M., (2006). Data Mining: Concepts and Techniques. San Francisco: Morgan Kaufmann Publishers.