Web Content Mining: A Solution to Consumer's Product Hunt

Syed Salman Ahmed, Zahid Halim, Rauf Baig, and Shariq Bashir

***Abstract*—**With the rapid growth in business size, today's businesses orient towards electronic technologies. Amazon.com and e-bay.com are some of the major stakeholders in this regard. Unfortunately the enormous size and hugely unstructured data on the web, even for a single commodity, has become a cause of ambiguity for consumers. Extracting valuable information from such an ever-increasing data is an extremely tedious task and is fast becoming critical towards the success of businesses. Web content mining can play a major role in solving these issues. It involves using efficient algorithmic techniques to search and retrieve the desired information from a seemingly impossible to search unstructured data on the Internet. Application of web content mining can be very encouraging in the areas of Customer Relations Modeling, billing records, logistics investigations, product cataloguing and quality management. In this paper we present a review of some very interesting, efficient yet implementable techniques from the field of web content mining and study their impact in the area specific to business user needs focusing both on the customer as well as the producer. The techniques we would be reviewing include, mining by developing a knowledge-base repository of the domain, iterative refinement of user queries for personalized search, using a graph-based approach for the development of a web-crawler and filtering information for personalized search using website captions. These techniques have been analyzed and compared on the basis of their execution time and relevance of the result they produced against a particular search.

***Keywords*—**Data mining, web mining, search engines, knowledge discovery.

I. Introduction

COMPUTERS have made our lives easier, but the Internet has made it really difficult. Using computers today we can solve the most difficult and complicated of problems; Internet was designed to share knowledge to help solve some of the mysteries of computer science. Internet is also a great means of doing business today. Today the Internet is being used for a lot many purposes apart from the ones stated above. But it has a problem, which could not have been predicted at the time of its inception.

The Internet is a huge collection of data that is highly unstructured which makes it extremely difficult to search and retrieve valuable information. The size of Internet is growing exponentially and there is no force monitoring or handling its contents to structure it. Any sort of computation being performed by a computer needs data to be in some specified order, without which it cannot perform its assigned task (unlike human brain which fascinatingly can process a lot of unstructured information with extensive ease).

The biggest problem that a shopper faces today is the number of *"uninteresting documents"* that are returned to him as a result of a simple search, which result in a lot of time wastage, in browsing through useless links.Although present day's web searching capabilities, networking and computational efficiency has allowed the user with huge bandwidth and very fast downloading speeds, the time wasted in browsing through the uninteresting documents is enormous.

When we talk of data we speak of not only text but also video and audio type data as well.

So how does one manage such data? The solution lies in Web content mining. *Search engines define content by keywords. Finding contents' keywords and finding the relationship between a Web page's content and a user's query content is content mining.* [1]

In this paper we review some of the latest techniques that are being used in the line of web content mining. The rest of the paper is organized as: section 2 deals with mining by developing a knowledge-base repository of the domain, section 3 presents iterative refinement of user queries for personalized search, section 4 describes using a graph-based approach for the development of a web-crawler, section 5 explains filtering information for personalized search using website captions, section 6 shows the use of web-based languages like XML semantics and conclusion concludes the paper.

II. Mining by Developing a Knowledge-Base Repository of the Domain

When information is given a well-defined meaning by defining the relationship between web pages and their contents, we are said to be creating an ontology. However this definition of relationship is very difficult to identify for a number of reasons, some of which are,

i. The identification of *vocabulary* that is used to describe the relevant concepts within the document.
ii. Finding *definitions*, for the vocabulary identified, that best describes the term.
iii. Identifying *correct relations* between the above two, where one term may be linked to many definitions and one definition may be for more than one term.

Authors are with Department of Computer Science, National University of Computer and Emerging Science, FAST-NU H11/4 Islamabad, Pakistan (e-mails: ssalman99@gmail.com, zahid.halim@nu.edu.pk, rauf.baig@nu.edu.pk, shariq.bashir@nu.edu.pk; http://www.nu.edu.pk, http://www.ming.org.pk/).

There are, however, methods available that can be used as solution for the above mentioned issues and can result in the formation of a very good ontology. One of such methods is the construction of an active knowledge base.

A. Build an Ontology

A knowledge base may be build by collecting information from a group of domain experts. This stage involves both time and money, so an easier approach may be to begin by constructing the ontology for the desired domain only. The goal of this ontology must be to reduce the conceptual and terminological confusion among members of the community and can be achieved by identifying and defining a set of relevant concepts characterizing an application domain.

Structure of the Ontology

According to Navigli [2], we can characterize an ontology into a formulated structure having the following three main components.

i. Top Ontology

These form the basic principles on the basis of which the entire ontology will be built.

ii. Upper Domain Ontology

It is a collection of key domain conceptualizations.

iii. Specific Domain Ontology

This level provides the details of the domain identified previously.

Features Necessary for a Usable Ontology

In the light of the structure of the domain ontology just explained, Navigli [2] has also identified the following 3 key features necessary in the construction of a usable ontology.

Coverage

This requires that the specific domain must be sufficiently populated so as to provide with the desired information. Under populating this level can mean failure to the system. However at this stage a trade-off needs to be struck between the cost to build such a system and its correctness.

1. Consensus

There must be a unanimous agreement on the business rules the system is working on between the domain experts building and using the system.

Accessibility

The ontology built should be implementable and easily accessible to users.

Enrich Ontology

An important factor to getting a useful ontology is the maintenance of the ontology by enriching it with additional features apart from just the definitions. This can be achieved by the following three methods as described by Navigli [2].

i. User Hyponymy Patterns

Gather relations between documents and their content in the form of a hyponym relationship. This means that forming relations like: *Aristotle, the philosopher*. Such relations, though difficult to find and implement, work well in complementing the basic ontology.

ii. Use Domain Terms

This involves the use of statistical methods and string inclusion to create syntactic trees.

iii. Use Statistical Classifiers

A classifier is used to associate and find semantic roles automatically, on the basis of gathered statistics, between the terms.

iv. Use Machine Learning Tools for creating Ontology

Collecting information from the web and incorporating it into the current ontology is vital for the freshness of the system. This can be achieved by using machine-learning tools to automatically gather information from the web and use it to build on the existing ontology.

B. Extraction of Terminologies for Domain Construction

From the domain corpus, the candidate terms that are necessary in the construction of the ontology can be gathered. That term is interpreted by assigning it proper definition. Here multiple definitions may be assigned. Then the relationship between the term under consideration and the existing terms is identified and ranked links are developed between similar terms. This process is refined with the three sub-sections described above. As a result of this process the ontology derived is representative of the specific domain only.

A diagrammatic representation of the extraction of terms for the constitution of a domain ontology is presented in Fig. 1.

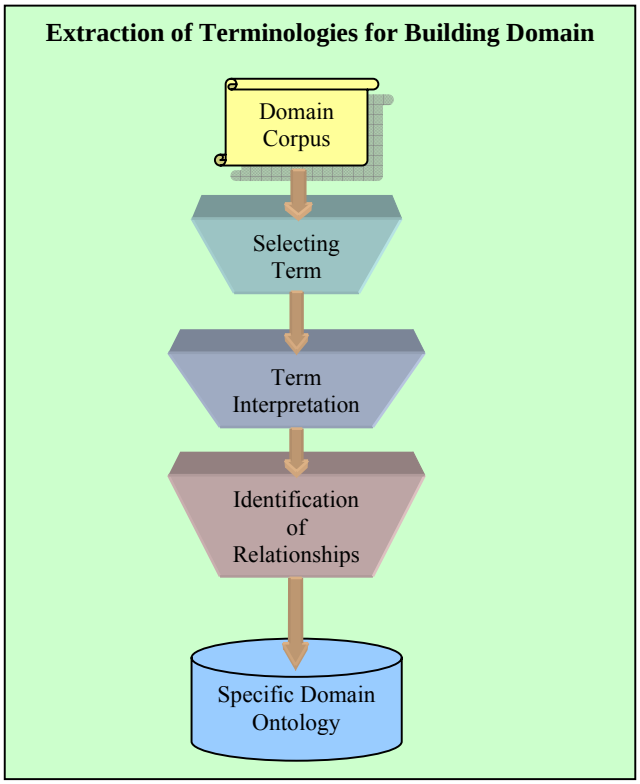

Fig. 1 Extracting Terminologies for Domain

III. Iterative Refinement of User Queries for Personalized Search

This is a very interesting approach, where the initial results of the user's query are re-ordered, by observing user preferences, and by asking some questions to assist the user in the search process.

One might argue with the need for this approach with the existence of powerful search engines available to us today like Yahoo! © Well just try and look up a keyword on it. If one compares the total number of hits resulted by it with the actual desired ones, he will realize the importance of query refinement search engines, such as this technique proposes. Imagine the uses of such a technique just in shopping and the time it would save a conventional customer!

A fresh induction to this area of personalized search is a Multiplicative Adaptive algorithm. This method uses an iterative query technique to refine user search by learning about user preferences.

A. Query Reformulation Methods

The adaptive expansion of the query reformulation can take place by either of these two methods as defined by Meng in [3]:

i. Linear

The updation is done linearly. i.e.

$$f(x) = \alpha x$$

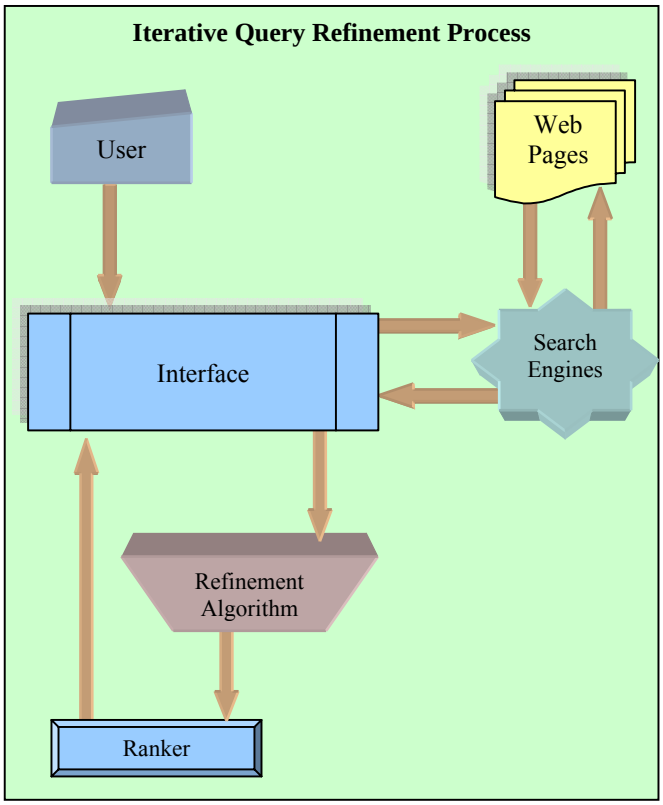

Fig. 2 Iterative Query Refinement Process

ii. Exponential

The updation is done exponentially. i.e.

$$f(x) = \alpha^{x}$$

B. Process of Iterative Query Refinement

The queries in this method are accepted through an interface, passed to any general purpose search engine available. The search engine would retrieve the entire list of web pages that according to the search criteria of the search engine appear the best. This list is passed on to the user interface once again, where by the user can choose the best desired links at random from the initial list. This list is passed on to the Refinement Algorithm, which adjusts the weights of the pages resulted from the search. The Ranker finally assigns new ranks to the pages and the new list is displayed before the user. The user may choose to further fine tune the results displayed, or if the results are rather to his liking he may accept the results and continue with the traversal of the pages to carry on with the search.

Fig. 2 is an elaborate diagrammatic representation of the procedure described above.

C. Result Confirmation

In order to confirm upon the success of the refined pages retrieved, the following statistical methods can be used.

1) Precision

Used to assess the performance of the algorithm and compare it with that of the meta-search engine. It is the ratio between the number of relevant documents returned and the total number of relevant documents returned according to Meng [3].

$$P_r = \frac{|R_m|}{M}$$

2) Relative Placement of relevant results

An Average Rank L_m, of the relevant documents in a returned set of m documents is defined in [3] by Meng as:

$$L_m = \frac{\sum_{i=1}^{C_m} L_i}{C_m}$$

where L_i is the rank of a relevant document i among the *top-m* returned documents. C_m is the count of relevant documents among the m returned documents.

3) Time Taken

The time taken by the entire process is the sum of the initial time taken by the general purpose search engines plus the time taken by the refinement algorithm to process the results.

IV. Using a Graph-Based Approach for the Development of a Web-Crawler

This is another very important and interesting method used in web content mining. It involves the construction of a focused web-crawler *[4]* (a goal-oriented web robot, aiming at retrieving only the subset of the web that is related to the given topics.

But building a focused web-crawler is not an easy task. The difficulty underlying in building the crawler is the assignment of proper order to web pages. Where some documents have higher priority for one customer, they may have none for another, yet the traditional search engines pop may pop them for viewing before the user irrespective of their use.

So the solution to the above-mentioned problem is the construction of a system that establishes the relevancy between pages and their keywords.

One method of finding relevancy between topics and documents to be retrieved is the use of Relevancy–Context Graphs. In this approach the only thing the user need show the system is the topic.

Using Relevancy Context Graphs involves the construction of a topic-specific web search engine. The search engine would use domain knowledge and focused web crawlers to facilitated search. Ranks assigned to pages can also facilitate the search process (method adopted by Google ™). Categorization of textual information by using supervised or unsupervised learning classifiers can also be a part of this system. Finally some documents that are not required may lead to highly relevant documents. Storing knowledge of the link hierarchies, by using relevancy-context graphs, can cater for this issue.

Relevancy Context Graph

For the rest of the discussion in this section, we will focus on the working of the relevancy context graphs.

It has been observed that hyperlinks of a page always provide semantic linkages between pages, also most pages are linked by other pages with related contents and that similar pages have links pointing to related pages [4]. Based on this assumption, a relevancy context graph is constructed for each on-topic document (topic about which query was made). The links of the graph have semantic relationships based on the assumption of topic locality. The documents that are the most closely related are placed in the inner most layer, the ones less related than the first layer are placed away, maybe in the second layer depending on their relative closeness with the other documents and so on.

To model this phenomenon, a number α is used, ranging from 0 to 1, to represent the relationship between documents and the desired query. When the power of α is 0, the document is on-topic or an exact match with the given query, when power of α is 1 it means that some dissimilarity between the document and the query exists; and so on. Fig. 3 below shows a relevancy context graph.

The relevancy of a document may be measured by a number of factors including the website caption, keywords in the document, page rank or the hyperlinks in the documents pointing to other useful links

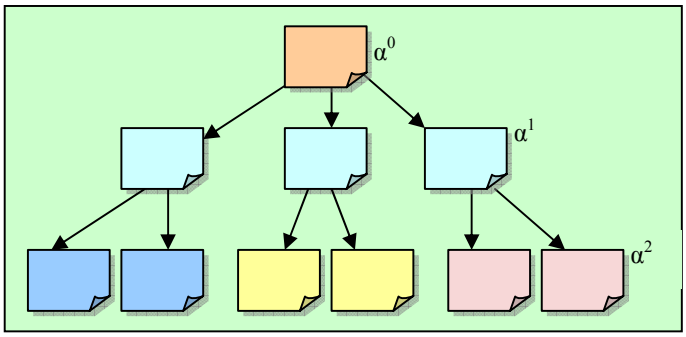

Fig. 3 Context graph with relevancy judgment

V. Filtering Information for Personalized Search Using Website Captions

A major part of the web is non-text data (images, video, audio, softwares); searching such data is extremely difficult, because its contents are unknown until studied, at least to some extent. Analysis of the captions for such data is an efficient method of collecting information without actually going through the entire data. Captions are text that describe both text and non-text data and are used for the convenience of the user. [5]

The problem with mining using captions is that searching them and understanding them is a huge issue. Searching web-captions is difficult because they may not be clearly or explicitly identified. Understanding them is difficult because of the wide number of varying semantics available on the web to map captions on a page. Some of the giants in the field of web content mining using captions (Yahoo!) face difficulties in indexing non-text data. Often the data retrieved is not of the highest importance to the user as it should have been.

A. Methods of Getting and Analyzing Captions from the Web

i. Syntactic Method

A web page is coded as tagged information; captions can be obtained from these tags.

ii. Semantic Method

After captions have been derived using the syntactic method, their grammar is studied which gives useful insight into the type of data the caption is referring to.

iii. Mathematical Method

These methods quantify the appropriateness of caption to the search being conducted, by comparing their strengths.

VI. Conclusion

The results generated by any general purpose search engine do not necessarily produce result that is the best possible according to user need. To improve on this issue one method of searching content on web without going through the actual content is web content mining. The implementations of these techniques lead to significant time improvement, which is very important factor for any business user.

REFERENCES

[1] Laware, G. W., Metadata management: A requirement for web warehousing and knowledge management. (2005) Scime, A. Web mining: Applications and Techniques, PA: Idea Group.
[2] Navigli, R., Ontology Learning from a Domain Web Corpus. In: Scime, A. (Ed.), Web Mining: Applications and Techniques, Idea, London. pp. 69-98.
[3] Chen, Z. - Meng, X., MARS: Multiplicative Adaptive Refinement Web Search. In: Scime, A. (Ed.), Web Mining: Applications and Techniques, Idea, London. pp. 99-118.
[4] Wu, F. and Hsu, C., Using context information to build a topic specific crawling system. In: Scime, A. (Ed.), Web Mining: Applications and Techniques, Idea, London. pp. 50-68.
[5] Kotb, Y., Gondow, K., Katayama, T., XML Semantics. In: Scime, A. (Ed.), Web Mining: Applications and Techniques, Idea, London. pp. 169-188.
[6] en.wikipedia.org/wiki/Web_mining.html
[7] http://www.eg.bucknell.edu/~xmeng/mars/mars. html