ABSTRACT

Web page clustering techniques categorize & organize search results into semantically meaningful clusters that assist users to search relevant information quickly. In general, it provides a solution for data management, information locating & interpretation of web data. Also facilitate users for discrimination, navigation & organization of web pages. Finding information on the World Wide Web is one of the most popular activities of Internet users. Due to increasing amount of information on the web, it has become very important to organize this large amount of information into meaningful clusters. Clustering is currently one of the most crucial techniques for dealing, with massive amount of heterogeneous information on the web. Unlike clustering in other fields, web page clustering separates unrelated pages and clusters related pages of a specific topic into one group. This paper gives an idea about Web Page rank algorithms like HITS (Hyperlink Induced Topic Search); page Rank, & Web page clustering algorithms like STC (Suffix Tree Clustering), Vivisimo Algorithm, Lingo algorithm with their advantages & Disadvantages.

Categories and Subject Descriptors
H.3.3 [Information storage and retrieval]: Information search & retrieval, Clustering. H.3.53 [Information storage and retrieval]: Online Information Services-Web-based services
General Terms
Algorithm, Experimentation, Design
Keywords

1. INTRODUCTION

clustering of data is a method by which large sets of data are grouped into clusters of smaller sets of similar data. Clustering can be considered the most important

Unsupervised learning problem.
Current web searches are possible because of search engines, such as Yahoo, AltaVista and Google. In most cases, however, the returned search results are also a large information source from which it is still difficult for users to find required information. This is called as web search problem. To solve this problem special kind of web service- search engines were created which solves this problem up to some extent but still far from perfect solution. How to effectively and efficiently manage and retrieve web-searched information is becoming a challenging problem. The exponential growth of information on the World Wide Web has prompted for developing efficient and effective methods for organizing and retrieving the information. Clustering techniques play an important role in searching and organization of web pages.

Traditional IR approaches are hardly appropriate in the context of the web, due to both the enormous size and hyper linked nature of the web. The majority of search engines give a long list of ranked web pages; most of them are irrelevant. The low precision of the web search engines coupled with the long ranked list presentation make it hard for users to find the information they are looking for. Typical queries retrieve hundreds of documents, most of which have no relation with what the user was looking for.

The limitations of search technology can be attributed to the following:

Polysemy: The words involved in the search have multiple meanings. For example, a user searching for windows may be interested in either the operating system or the physical object.

Phrases: a phrase may be different from words in it. e.g., the meaning of the phrase “Partition magic” (a disk partition management tool) is quite different from the meaning of the individual words “partition” and “magic”.
Term dependency: words in the terms are not totally independent of each other. For example, a user may look for details about a product made by a particular company and type in Sun’s Enterprise Computer Series. Obviously, each word in this term is dependent of each other.

One possible solution to improve the quality of search is a new approach, called web page clustering was formulated. Instead of showing the user a long list of results itself, it is split into groups related to sub-topics. If the user is shown these groups, possibly with some keyword type descriptions, they can then select one (or more) that fit their perceived interests. Clustering algorithms attempt to group web pages together based on their similarities; thus pages relating to a certain topic will hopefully be placed in a single cluster. This can help users both in locating interesting web pages more easily and in getting an overview of the retrieved document set. Several researchers have suggested that the clustering techniques are feasible for web mining.

2. HOW DO SEARCH ENGINES WORK

Search Engines like Google, Yahoo, and Ask searches a database of the full text of web pages automatically from the billions of web pages that are residing on servers. Search engine databases are selected and built by computer robot programs called spiders. These crawl the web, finding pages for potential inclusion by following the links in the pages they already have in their database. If a web page is never linked to in any other page, search engine spiders cannot find it. After spiders find pages, they pass them on to another computer program for indexing. This program identifies the text, links, and other content in the page and stores it in the search engine database's files so that the database can be searched by keyword and whatever more advanced approaches are offered, and the page will be found if your search matches its content. The general architecture of web search engine is as shown in fig. (a)

3. KEY REQUIREMENTS FOR WEB PAGE CLUSTERING

As pointed out by Zamir and Etzioni [18] the followings are the key requirements for web page clustering methods.

3.1. Relevance: The method must produce clusters that group web pages relevant to the user’s query.

3.2. Browsable Summaries: The user needs to determine at a glance whether a Cluster’s contents are of interest. Ranked lists of the clusters may infact difficult to browse. Therefore the method has to provide concise and accurate descriptions of the clusters.

3.3. Overlap: Since web pages have multiple topics, it may be member of more than one cluster, so it is important to avoid confining each web page to only one cluster.

3.4. Snippet-tolerance: The method must produce high quality clusters even when it only has access to the snippets returned by the search engines, as most users are unwilling to wait while the system downloads the original documents off the Web.

3.5. Speed: A very patient user might go through 100 documents in a ranked list presentation. Clustering on the other hand allows the user to browse several related documents. Therefore the clustering method ought to be able to cluster up to one thousand snippets in a few seconds.

4. BACKGROUND & MOTIVATION

Due to the exponential growth of information on the World Wide Web, web search is difficult for user. Typical queries retrieve hundreds of web pages, most of which have no relation with what the user was looking for & unable to find required information efficiently. Therefore, improving the performance of web page clustering algorithm is worthwhile & very important.

4.1. Why web page clustering

As the size of stored data grows, it is getting harder and harder to find the relevant information that one is looking for. Web search is difficult because locating truly needed web pages & interpreting them appropriately is a big challenge faced by researchers in the fields of database, Information Retrieval (IR) & data mining. It is difficult for users to write queries that are descriptive & discriminating to find just the web pages that user required.

Web page clustering is one approach for assisting users to comprehend the result set & to refine the query. It is useful for formation of meaningful groups of web pages & presents these to the user as clusters.

Clustering performance is very important, if cluster quality is poor; cluster will be semantically meaningless or will contain many irrelevant pages. Therefore, for improving the performance of web page clustering algorithm’s cluster quality should be improved.

5. WEB PAGE RANKING ALGORITHMS
Ranking web pages using the information contained in the hyperlinks between web pages currently is an active research area. Among the existing page ranking algorithms the most important algorithms are Kleinberg’s HITS algorithm [14], Brin & Page’s Page Rank algorithm [19]. The AltaVista Search Engine implements HITS algorithm. But the HITS (Hyperlink Induced Topic Search) is a purely link structure-based computation, ignoring the textual content [1]. According to Page Rank algorithm used in Google [2], a page has a high rank if the sum of the ranks of its back-links is high.

5.1.1. HITS Algorithm

It is a query-based algorithm. Given a user query, the HITS algorithm first creates a neighborhood graph for the query. The neighborhood contained top 200 matched web pages retrieved from a content-based web search engine; it also contains all the pages of the 200 web pages linked to and pages that linked to these 200 top pages. Then, an iterative calculation was performed on the value of authority and value of hub. For each page p, the authority and hub values are computed. The authority value of page p is the sum of hub scores of all the pages that points to p, the hub value of page p is the sum of authority scores of all the pages that p points to (Fig.1). Iteration proceeded on the neighborhood graph until the values converged. Then, an iterative calculation was performed on the value of authority and value of hub.

For each page p, the authority and hub values are computed as follows:

\[
a_i = \sum_{j \in B(i)} h_j \quad \text{and} \quad h_j = \sum_{i \in F(j)} a_i.
\]

5.1.2. Page Rank Algorithm

The Page Rank algorithm is used by Google search engine. Unlike HITS algorithm, Page Rank is a query-independent algorithm, which is, assigning a rank score to each page independent of a given query; it is based on the connectivity structure of the web pages. The Page Rank value of a page is weighted by each hyperlink to the page proportionally to the quality of the page containing the hyperlinks; i.e., the Page Rank value of a page will spread evenly to all the pages it points to. As shown in Fig.2, page q1 has two hyperlinks points to page p and page m separately, thus, the Page Rank value of page q1 will be distributed to p and m evenly, and each of them gets 50 points. Since page p has two link points to it, the Page Rank value of page p is the sum of weights of the incoming links.

6. WEB PAGE CLUSTERING ALGORITHMS

Researchers have applied the entire standard clustering methods [2,8,15] to web page clustering like Hierarchical (agglomerative and divisive), Partitioning (Probabilistic, k-medoids, K-means), and many more. Many algorithms build on the standard methods by using web or document specific are: Suffix Tree Clustering (STC)[18] and Lingo [10, 11], Extended suffix Tree clustering (ESTC), Vivisimo [17, 9].

6.1.1. STC Algorithm

STC is a linear time clustering algorithm that is based on a suffix tree which efficiently identifies sets of documents that share common phrases. STC treats a document as a string, making use of proximity information between words. STC is incremental and O (n) time algorithm. STC has three logical steps: (1) document cleaning, (2) identifying base clusters using a Suffix tree, and (3) combining these base clusters into clusters.

Step 1 - Document Cleaning

In this step, the string of text representing each document is transformed using a light Stemming algorithm (delete...
word prefixes & suffixes and reducing plural to singular). Sentence boundaries (identified via punctuation and HTML tags) are marked and nonword tokens such as numbers, HTML tags and most punctuation are removed.

**Step 2 - Identifying Base Clusters**
The identification of base clusters can be viewed as the creation of an inverted index of phrases for the document collection. This is done efficiently using a data structure called a suffix tree. This structure can be constructed in time linear with the size of the collection, and can be constructed incrementally as the documents are being read. Each node of the suffix tree represents a group of documents and a phrase that is common to all of them. Therefore, each node represents a base cluster. Furthermore, all possible base clusters (containing 2 or more documents) appear as nodes in our suffix tree. Each base cluster is assigned a score that is a function of the number of documents it contains, and the words that make up its phrase.

**Step 3 - Combining Base Clusters**
Documents may share more than one phrase. As a result, the document sets of distinct base clusters may overlap and may even be identical. To avoid the proliferation of nearly identical clusters, the third step of the algorithm merges base clusters with a high overlap in their document sets (phrases are not considered in this step). The STC algorithm is incremental and order independent. As each document arrives from the Web, we "clean" it and add it to the suffix tree. Each node that is updated (or created) as a result of this is tagged. We then update the relevant base clusters and recalculate the similarity of these base clusters to the rest of the base clusters. If there is any changes in the base cluster graph result in any changes to the final clusters. The final clusters are scored and sorted based on the scores of their base clusters and their overlap. As the final number of clusters can vary, the top few clusters need to be reported. Typically, only the top 10 clusters clusters of interest. For each cluster reported the number of documents it contains, and the phrases of its base clusters. In STC, as documents may share more than one phrase with other documents, each document might appear in a number of base clusters. Therefore a document can appear in more than one cluster. Note that the overlap between clusters cannot be too high; otherwise they would have been merged into a single cluster.

6.1.2. **Vivisimo Algorithm**
It uses a specially developed heuristic algorithm to group - or cluster - textual documents. This algorithm is based on an old artificial intelligence idea: a good cluster - or document grouping - is one, which possesses a good, readable description. So, rather than form clusters and then figure out how to describe them, they only form well-described clusters in the first place [14]. Vivísimo is doing hierarchical, document clustering, conceptual, on-the-fly techniques.

6.1.3 **Lingo Algorithm**
It is able to capture thematic threads in a search result, that is discover groups of related documents and describe the subject of these groups in a way meaningful to a human.

6.1.3.1 **Preprocessing**
First step is stemming and stop words removal, current experiments show that preprocessing is of great importance in Lingo because the input snippets are automatically generated summaries of the original documents and hence are usually very small (one or two sentences). In this algorithm for stemming documents, Porter’s algorithm is used.

6.1.3.2 **Frequent phrase extraction**
We define frequent phrases as recurring ordered sequences of terms appearing in the input documents. Intuitively, when writing about something, we usually repeat the subject-related keywords to keep a reader’s attention. Obviously, in a good writing style it is common to use synonymy and pronouns and thus avoid annoying repetition. The suffix arrays approach is used, because it is convenient in implementation and very efficient.

6.1.3.3 **Cluster label induction**
Once frequent phrases (and single frequent terms) that exceed term frequency thresholds are known, they are used for cluster label induction. There are three steps to
this: term-document matrix building, abstract concept discovery, Phrase matching and label pruning.

6.1.3.4 Cluster content discovery
In the cluster content discovery phase, the classic Vector Space Model is used to assign the input documents to the cluster labels induced in the previous phase.

6.1.3.5 Final cluster formation
Finally, clusters are sorted for display based on their score, calculated using the following simple formula:
\[ C_{score} = \text{label score} \times k_C \]
Where \( k_C \) is the number of documents assigned to cluster \( C \). The scoring function, although simple, prefers well-described and relatively large groups over smaller, possibly noisy ones. For the time being, no cluster merging strategy or hierarchy induction is proposed for Lingo.

7 Limitations of current web page clustering algorithms
These algorithms produce clustering of low quality.

Producing semantically meaningless clusters. Semantically meaningful clusters are often small, missing many relevant pages and contain irrelevant pages [16]. The problem is that these algorithms only use textual properties and static’s of pages from the result set. Other algorithms such as partitioning and hierarchical algorithm [15] use data similarity measures [2] to construct clusters. But these similarity-based methods are not effective to produce semantically meaningful clusters as these are directly applied to page data.

One way of improving web page clustering algorithms is to make better use of the textual properties of web pages. The semantic relationships between words is very useful information; for example, synonyms, hyponyms, meronyms, etc. [4]. Word Net [4] is a lexical reference system and is one source of this information. However, the data in these systems is incomplete, particularly for commercial, technical, and popular culture word usage.

An alternate source, although less accurate and less informative, is to use global document analysis and term co occurrence statistics to identify whether terms are related or unrelated. The number of pages in multi-term search result sets can approximate term co-occurrence statistics.

Providing the most relevant pages earlier in the results can reduce the time users spend searching [1]. Most clustering algorithms order the pages in the clusters by their position in the search results [3]. Such an ordering fails to use the additional information about the user’s search goal, provided by the user selecting the cluster, so the most relevant pages may not be shown first.

8 CONCLUSION
Clustering is not a brand-new technique. This technique has been used in the statistics for last five decades. The IR community has explored web page clustering as an alternative method of organizing retrieval results, but clustering has yet to be deployed on most major search engines. Industry analysts predict that Google and other major search engines will need to make use of clustering technology to stay competitive.

In this paper, we have done analysis of current web page ranking algorithms like HITS, page rank, & web page clustering algorithms like STC, Vivisimo Algorithm & Lingo algorithm with their advantages & Disadvantages.

REFERENCES


