A Survey on Web Clustering Engine

Bhavana R.potrajwar
CSE,SGBA University
bhavanapotrajwar@gmail.com

Abstract- **World Wide Web is a very large distributed digital information space. The ability to search and retrieve information from the Web efficiently and effectively is an enabling technology for realizing its full potential. Current search tools retrieve too many documents, of which only a small fraction are relevant to the user query. Web clustering engine greatly simplifies the effort of the user from browsing e large set of search results by reorganizing them into smaller clusters. It organizes search results by topic, thus offering a complementary view to the flat-ranked list returned by conventional search engines. This paper highlight the main characteristics of a number of existing Web clustering engines and also discuss how to evaluate their retrieval performance.**

I. INTRODUCTION

There is lot of information on the World Wide Web, this information are available in unstructured, disorganized, dynamic and heterogeneous in nature and enormously large.It has become difficult to desire information on search engine.By using clustering techniques grouping similar documents together in order to facilitate presentation of results in more compact form and enable thematic browsing of the results set ,this approach solve the problem of information retrieval. The four main criteria for creating cluster categories: Making the titles concise, accurate, distinctive, and "humanlike" -- in other words, not something that looks like it was generated by a machine. More specifically, it is a process of grouping similar documents into clusters so that documents of one cluster are different from the documents of other clusters. There are many web clustering engines available on the web (Carrot2,Vivisimo, SnakeT, Grouper etc.) which give the search results in forms of clusters.[2]. A web clustering engine takes the result, returned by the search engine as input and performs clustering and labeling on that result. One common feature of most current clustering engines is that they do not maintain their own index of documents; similar to meta search engines, they take the search results from one or more publicly accessible search engines. 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. It takes lot of time to find the relevant information. Typical queries retrieve hundreds of documents, most of which have no relation with what the user was looking for. According to this, we considered Web-snippet clustering engine is a useful complement to the flat, ranked list of results offered by classical search engines (like Google).Web snippet (short description) clustering also known as Web Search Results Clustering is an attempt to apply the idea of clustering to snippets returned by a search engine in response to query. A clustering engine tries to address the limitations of current search engines by providing clustered results as an added feature to their standard user interface and meaningful labels.

II. WEB CLUSTERING ENGINES GOAL AND ARCHITECTURE

A. Web Clustering Engines

Plain search engines are usually quite effective for certain types of search tasks,such as navigational queries (where the user has a particular URL to find) and transactional queries (where the user is interested in some Web-mediated activity).However, they can fail in addressing informational queries (in which the user has an information need to satisfy), which account for the majority of Web searches . This is especially true for informational searches expressed by vague, broad or ambiguous queries. This is especially true for informational searches expressed by vague, broad or ambiguous queries.[1] A clustering engine tries to address the limitations of current search engines by providing clustered results as an added feature to their standard user interface. We emphasize that clustering engines are usually seen as complementary—rather than alternative—to search engines. In fact, in most clustering engines the categories created by the system are kept separated from the plain result list, and users are allowed to use the list in the first place. The view that clustering engines are primarily helpful when search engines fail is also supported by some recent experimental studies of Web searches .The search aspects where clustering engines can be most useful in complementing the output of plain search engines are the following.

—Fast subtopic retrieval. If the documents that pertain to the same subtopic have been
correctly placed within the same cluster and the user is able to choose the right path
from the cluster label, such documents can be accessed in logarithmic rather than
linear time.
—Topic exploration. A cluster hierarchy provides a high-level view of the whole query
topic including terms for query reformulation, which is particularly useful for informational
searches in unknown or dynamic domains.
—Alleviating information overlook. Web searchers typically view only the first result
page, thus overlooking most information. As a clustering engine summarizes the

content of many search results in one single view on the first result page, the user

may review hundreds of potentially relevant results without the need to download and scroll to subsequent pages.

B . SEARCH ENGINE

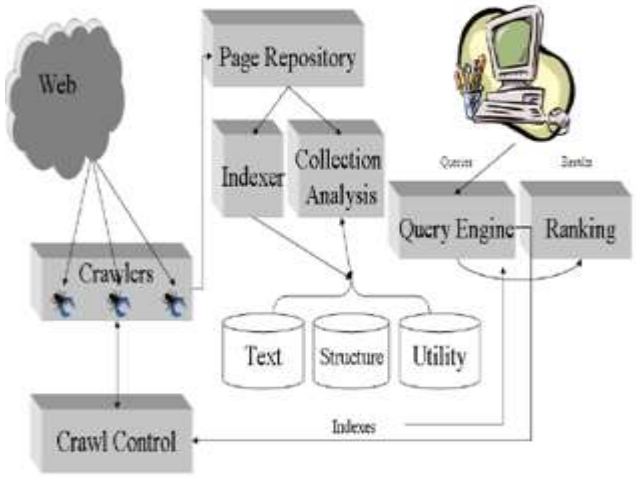

Fig 1. Architecture of Web Search Engine

The Search Engine component is a part of the Information Retrieval model component. Its main responsibility is the comparison of documents based on their document models through obtaining documents similarity values. In today's search engines, Clustering of results is the next step up from ranking of documents Web search engines did not come into existence until 1994.A search engine has four components:

- document processor indexes new documents. Indices are a mapping between words and what documents they appear in. Most engines are spider-based, so a crawl of the web for new documents and the updating of the index is automated.
- query processor inspects a user's query and translates it into something internally meaningful.
- matching function uses the above internally meaningful representation to extract documents from the index.
- ranking scheme positions the more-relevant documents on top, using some relevance measure.

We may classify web search engines according to the set of features they explore (Broder, 2002). First generation web search engines, starting in 1994 with WebCrawler and Lycos, explore on-page data (content and formatting). They support mostly informational queries. The second generation, emerging in 1998, with Google (Page Rank), uses off-page web specific data (link analysis, anchor text and click streams data) and supports both informational and navigational queries. The third generation, appearing during the first years of 2000 attempts to merge multiple sources of evidence and aims to support all kinds of queries. Modern Web IR is a discipline which has exploited some of the classical results of Information Retrieval developing innovative models of information access. Therefore, search engines have established as a revolutionary working metaphor. If someone needs information about a book, an address, a research paper, a flight ticket, or almost any other topic, they just make a query on a search engine. In this paragraph we briefly review the architecture of a typical search engine. The architecture of a search engine is given in Figure 3.Crawlers are distributed agents which gather information from the Web. They crawl the Web graph visiting the pages according to some policies (BFS, DFS, random, topic focused, prioritized) and store the pages in a local Page Repository. From time to time, the pages are indexed and analyzed to produce a fast indexable representation of both the documents and the link structures. Both the textual and structural indexes are then analyzed to rank the documents stored in the repository. For efficiency reasons, part of this ranking process can be performed off line, before the query is submitted through the query engine. Nowadays, modern search engines index billions of objects on distributed platforms of thousands of commodity PCs running Linux. The index is often organized in a two tier structure and is replicated by thousands of autonomous clusters of servers. This distribution is needed to sustain the peak load of thousands of queries per second. Users communicate with the query processor, which is the only visible component. It carries out several tasks, usually (but not limited to):

- tokenizing of the query to remove invalid characters, and to recognize meta-keywords or special syntactic operators.
- removal of stopwords; words which are too common and rarely help in the search (e.g. the, a, of, to, which).
- stemming; a process designed to improve the performance of IR systems, involving normalizing semantically similar words to their root forms (e.g. produce, produced, producer, producers, produces and producing map to produc-).
- assigning a weight to each keyword/keyphrase, to aid with ranking(Salton & Buckley 1988).

After results are retrieved by the matching function, they are ranked by relevance based on some ranking measure and set of heuristics (called the ranking scheme). Often taken into account are:

- term frequency how many times keywords appear in the document.
- inverted document frequency a value which aims to determine how important a term is in discriminating a document from others(Salton 1989).
- semantic proximity words synonymous to a given keyword may be matched, boosting the score of the document.
- term position keywords appearing in the title or heading (rather than the body) should contribute more to a document's weight.

- term proximity a document in which the query terms are close together is considered more relevant than one in which they are far apart.

- cluster distance how far apart groupings of matched terms are.
- percentage of query terms matched[3]

C. Web Clustering Engine

Name	Time complexity	Algorithm	clustering	Reference
Grouper	O(n)	STC	Flat	Zamir and Etzioni 1999
carrot	O(n)	Lingo	Flat	Weiss and Stefanowski 2003
Vivisimo	O(n2)		Hierarchical	
WICE	O(n)	SHOC	Hierarchical	Zhang and Dong 2004
Web Cat	O(nkt)	K means	Flat	F. Giannotti et al. 2003
SnakeT	O(n log n + m log mp)	Approximate Sentence Coverage	Hierarchical	Paolo Ferragina and Antonio Gulli 2005

Table 2 .comparisons of web clustering engine

- Grouper is a document clustering interface to the HuskySearch meta-search service and this based on MetaCrawer,retrives result from various search engine and Grouper clusters the results as they arrive using the STC algorithm.Grouper is start when user entering a query in query box.user can choose how query terms are treated and can specify the number of documents to be retrieved (10 – 200) from each of the participating search engines. If system required approximately 10 search engine then it retrives 70-100 documents after eliminating dublication. Main page diplay the number document retrives and number of cluster found. Cluster are present in table,and each cluster in a row is referred as summary.

- Carrot2 [Stefanowski, J. and Weiss, D (2003)] combines several search results clustering algorithms: STC,Lingo, TRSC, clustering based on swarm intelligence (ant-colonies), and simple agglomerative Techniques. Lingo uses SVD as the primary mechanism for cluster label induction. It is open source search result clustering engine.It automatically collect small documents. Carrot² offers ready-to-use components for fetching search results from various sources.Carrot² is written in Java and distributed under the BSD license.
- The first commercial clustering engine was probably Northern Light, at the end of the 1990s. It was based on a predefined set of categories, to which the search results were assigned. In this cluster and cluster lables are dynamically generated from search result. Vivísimo was founded by research computer scientists at the Computer Science Department at Carnegie Mellon University,where research was originally done under grants from the National Science Foundation. The company was founded in June 2000.It won the "best meta-search engine award" assigned by SearchEngineWatch.com from 2001 to 2003. It specially developed heuristic algorithm to group - or cluster - textual documents.

.

- WICE (Web information clustering engine) devise an algorithm called SHOC (semantic hierarchical online clustering) that handles data locality and successfully deals with large alphabets. For solving the first problem, SHOC [Dong 2002; Zhang and Dong 2004] uses suffix arrays instead of suffix trees for extracting frequent phrases.

- WebCAT [Giannotti et al. 2003], was built around an algorithm for clustering categorical data called transactional k-means. Originally developed for dadabase, it performs litile transaction. WebCAT's computational complexity is linear in the number of documents to be clustered, assuming a fixed number of iterations.

- SnakeT [Ferragina and Gulli 2004, 2005] is both the name of the system and the underlying algorithm. It interesting feature is builds a hierarchy of possibly overlapping folders. The computational complexity of the algorithm is O (nlog2 n + mlog2 mp), where n is the number of documents, m is the number of features, and p is the number of labels.

III. PERFORMANCE EVALUATION

Lancaster and Fayen (1973) once listed 6 criteria for assessing the performance of information retrieval systems.They are: 1) Coverage, 2) Recall, 3) Precision, 4) Response time, 5) User effort, and 6) Form of output.

Although the criteria were set up more than two decades ago and a great deal has been done to reduce user effort(e.g., design friendly user interface) in using the system, they still seem quite applicable to evaluating information retrieval systems today.

Based on our knowledge and experience gained from the current study, we believe that one needs to consider the following aspects when evaluating a Web search engine.

- Composition of Web Indexes

Whenever a Web search request is issued, it is the web index generated by Web robots or spiders, not the web pages themselves, that has been used for retrieving information. Therefore, the composition of Web indexes affects the performance of a Web search engine. There are three components that the authors would like to inspect regarding the makeup of a Web index, namely, coverage, update frequency and the portions of Web pages indexed (e.g., titles

plus the first several lines, or the entire Web page). We understand that the magnitude
of all three components depends largely on the power and sophistication of the hardware and software that make the Web index or database. On the other hand, larger coverage, frequent updates and fulltext indexing do not necessarily mean better Web search engines in other measurements.

- Search Capability

A competent Web search engine must include the fundamental search facilities that Internet users are familiar with, which include Boolean logic, phrase searching, truncation, and limiting facilities (e.g., limit by field).

- Retrieval Performance

Retrieval performance is traditionally evaluated on three parameters: precision, recall and response time. While the three variables can all be quantitatively measured, extra caution should be exercised when one judges the relevance of retrieved items and estimates the total number of documents relevant to a specific topic in the Web system.

- Output Option

This evaluation component should be examined from two perspectives. One is the number of output options a Web search engine offers, whereas the other deals with the actual content of the output. Sometimes, one search engine may appear quite impressive in one aspect, but in reality it cannot satisfy its users because of its weakness in the other facet of this evaluation criterion.

- User Effort

User effort refers to documentation and interface in this study. Well-prepared documentation and a user-friendly interface play a notable role in users' selection of Web search engines. Since there are more than two dozen of them available, the attractiveness of each Web search engine is expressed, to its users, mainly in its documentation and interface.[3]

IV. CONCLUSION

We have discuss goal of Web clustering engine and various existing Web clustering engine and there difference with each other on this Paper.By using clustering techniques ,user can find desired page fastely. To improve the search result clustering, First, more work needs to be done to improve the quality of the cluster labels and the coherence of the cluster structure. Second, the incrementality, because the web pages change very frequently and because new pages are always added to the web. Third, the fact that very often a web page relates to more than one subject should also be considered and lead to algorithms that allow for overlapping clusters. Fourth, Inconsistency is another problem. The contents of a cluster do not always correspond to the label and the navigation through the cluster sub hierarchies does not necessarily lead to more specific results. Fifth, advanced visualization techniques might be used to provide better overviews and guide the interaction with clustered results.

REFERENCES

[1] Broder, A. (2002) A taxonomy of web search. SIGIR Forum. 36:2. p. 3-10
[2] R. Subhashini et. al. / Indian Journal of Computer Science and Engineering Vol. 1
No. 4 392-401
[3] Carpenito C, Osinski S, Romano G, and Weiss D (2009) A Survey of Web Clustering Enginesý. ACM Computing Surveys, Vol. 41, No. 3, Article 17.
[4] Han J and Kamber M (2001) Data Mining -Concepts and Techniques. Academic Press.
[5] Kishwar Sadaf1 and Mansaf Alam2,International Journal of Computer Science & Engineering Survey (IJCSES) Vol.3,No.4, August 2012
[6] Zamir, O., Etzioni, O. 1998. Web document clustering: a feasibility demonstration.Proc.of SIGIR '98, Melbourne,Appendix-Questionnaire, pp.46-54
[7] Vivisimo.com. http://www.vivisimo.com
[8] Kishwar Sadaf1 and Mansaf Alam2,International Journal of Computer Science & Engineering Survey (IJCSES) Vol.3,No.4, August 2012

AUTHOR(S) PROFILE

Bhavana R.potrajwar received the B.E degree in Information Technology and pursing M.E. degree in CSE from JECT YAVATMAL,india,respectively.

