ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

A Survey Paper on Advanced Web Searching Using Lingo Algorithm

Amar Telange
JSPM'S Rajarshi Shahu CollegeOf Engineering, Tathawade Pune, Savitribai Phule Pune University, Pune.

Monika Patil
JSPM'S Rajarshi Shahu CollegeOf Engineering, Tathawade Pune, Savitribai Phule Pune University, Pune.

Amol Borse
JSPM'S Rajarshi Shahu CollegeOf Engineering, Tathawade Pune, Savitribai Phule Pune University, Pune.

Prof. Gitanjali.S.Mate
M.E Computer, Asst. Prof. IT Dept, JSPM'S Rajarshi Shahu College Of Engineering, Tathawade Pune, Savitribai Phule Pune University

ABSTRACT:

When the intention behind the search query is not clear, the search engine returns a large number of results. The results are displayed in the form of a ranked list. The aim of the search results clustering is to provide quick focus on relevant search results. The performance of the web search engines could be improved by properly clustering the search result documents. Ifwe effectively organize the web documents through the proper means of clustering techniques, we could definitely increase the performance of the search engines. Search results clustering problem is an automatic, on-line arrangementgrouping of similar documents in a search results list returned from a search engine. In this Report we present Lingo a novel algorithm for clustering search results, which emphasizes cluster description quality. We describe methods used in the algorithm: algebraic transformations of the term-document matrix and frequent phrase extractusing arrays. Finally, we discussed results acquired from an empirical evaluation of the algorithm for textual data.

KEYWORDS:

Information Retrieval,SearchEngines, Clustering, Lingo Algorithms Search Techniques.

INTRODUCTION:

The existing search engines always come out with a long list of results for the given query and they are ranked by their relevance to the same query. Information retrieval and ranking functions are vital to the search engines.

The organization and presentation of the results is also vital and could significantly affect the utility of the search engine [8]. Avast literature survey on page ranking and retrieval are being made by the researchers. But, there is relatively very little research that had been done on how to improve the effectiveness of search result organization [14]. The general concepts of the search engines are to focus upon the words that they find on a web page rather than the meaning of the words. In search result clustering, it is meant that the documents were returned in response to a query [3]. The default presentation of search results in information retrieval is a simple list. Users scan the list from top to bottom until they have found the information they are looking for. Instead, in the case of clusters similar documents appear together. It is often easier to scan a few coherent groups than many individual documents in disarray. This is particularly useful if a search term has different word senses.Clustering of web search results is an attempt to organize the results into a number of thematic groups in the manner a web directory does it [4]. This approach, however, differs from the human-made directories in many aspects. First of all, only documents that match the query are considered while building the topical groups [9]. Clustering is thus preformed after the documents matching the query are identified. Consequently, the set of thematic categories is not fixed – they are created dynamically depending on the actual documents found in the results. Secondly, as the clustering interface is part of a search engine, the assignment of documents to groups must be doneefficiently and on-line. For this reason it is difficult to download the full text of each document from the Web [14]. Clustering ought to be performed based solely on the snippets returned by the search service.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

REALATED WORK:

Clustering algorithms have been present in Information Retrieval for a long time, a comprehensive review of classic methods can be found in methods. One of their common applications was to organize large volumes of semi numerical data into entities of higher abstraction level, easier to perceive by humans. Agglomerative Hierarchical Clustering (AHC) and K-means algorithms gained widespread popularity when speed was of no such critical importance, because processing was performed online and only once in a while [10].These algorithms, used successfully in the economy, medicine or social sciences ware quickly transformed to the domain of search results clustering [4]. However, their computational complexity, difficult tuning of parameters and sensitivity to malicious input data soon raised the need of improvements.
Both of these discover phrases shared by document references in the search results and perform clustering according to this information [6].The former system introduces novel, Tree Clustering algorithm are produces at, but overlapping clusters, which is usually perceived as an advantage, because documents tend to belong to more than one subject. An extension of STC producing hierarchical structure of clusters was recently proposed in.Basically Lingo is used only for data which is generally considered as group of clusters. We research on several perspective areas which shows that lingo can be also implemented on images. Finally, clusters are sorted for display based on their score, calculated using the following simple formula: Score = label score × C, where 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.

EXISTING SYSTEM:

Basically, lingo framework is specially designed for textual data, but if user wants certain variations in framework like images, grouping of images and clusters, then it is difficult to predict what user actually wants. The previous system was consisting of data, which is form in one semantic group called as clusters. [5] Whenever user wants some of retrieval, he can give the input only in the form of text. In existing system the groups of clusters with same pages are compared, and then the actual result is elaborated.

Despite its drawbacks, we have decided to use the user-based evaluation method to assess the clustering results produced by LINGO [12]. Designing our algorithm we placed much emphasis on making the cluster labels accurate and meaningful to the end users. Therefore, we feel that is the users who can judge best to what extent LINGO achieved its goal. Below, we present the details on our proposed assessment scheme and report on the evaluation results.

PROPOESD SYSTEM:

The key characteristic of the Lingo Algorithm is that it first identifies cluster labels and then assigns documents to the labels to form final clusters [3]. To find the labels, Lingo builds a term document matrix for all input documents and decomposes the matrix to obtain a number of base vectors. Each such vector gives rise to one cluster label. To complete the clustering process, each label is assigned documents that contain the label's words.

Lingo identifies meaningful cluster labels using the Singular Value Decomposition (SVD), and then assigns documents to these labels to form proper clusters [12]. For this reason this algorithm could be considered as an example of a description-comes-first approach. The algorithm consists of five phases.

1. Preprocessing of the input snippets, which includes tokenization, stemming and stop-word marking.

2. It identifies words and sequences of words frequently appearing in the input snippets.

3. A matrix factorization is used to induce cluster labels.

4. Snippets are assigned to each of these labels to form proper clusters. The assignment is based on the Vector Space Model (VSM) and the cosine similarity between vectors representing the label and the snippets.

5. Post processing, which includes cluster merging and pruning?

The key component in label induction is an approximate matrix factorization, which is used to produce a low-dimensional basis for the column space of the term-document matrix.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

In linear algebra, base vectors of a linear space can be linearly combined to create any other vector belonging that space Therefore, in Lingo, each vector of the low-dimensional basis gives rise to one cluster label. The frequent word sequences or even single words appearing in the input snippets can also be expressed as vectors in the same vector space. Thus, the well-known measures of similarity between vectors, such as the cosine similarity, can be used to determine which frequent word sequence or single word best approximates the dominant verbal meaning of a base vector.

ALGORITHM: LINGO :

Input: A set of image documents D.
Output: A set of k clusters.
Method: 1: D input image documents.

{STEP 1: Preprocessing}

2: for all d D do
3: perform text segmentation of d; {Detect word boundaries etc.}
4: if language of d recognized then
5: apply stemming and mark stop-words in d;
6: end if
7: end for

{STEP 2: Frequent Phrase Extraction}

8: concatenate all documents;
9: Pc discover complete phrases;
10: Pf p : {p Pc frequency(p) > Term Frequency Threshold};

{STEP 3: Cluster Label Induction}

11: A term-document matrix of terms not marked as stop-words and
with frequency higher than the Term Frequency Threshold;
12: Σ,U, V SVD(A); {Product of SVD decomposition of A}
13: k 0; {Start with zero clusters}
14: n rank(A);
15: repeat
16: k k + 1;
17: q Σ ki=1 Σ ii)/(Σni=1 Σii);
18: until q < Candidate Label Threshold;
19: P phrase matrix for Pf ;
20: for all columns of UTk P do
21: find the largest component mi in the column;
22: add the corresponding phrase to the Cluster Label Candidates set;
23: labelScore mi;
24: end for
25: calculate cosine similarities between all pairs of candidate labels;
26: identify groups of labels that exceed the Label Similarity Threshold;
27: for all groups of similar labels do
28: select one label with the highest score;
29: end for

{STEP 4: Cluster Content Discovery}

30: for all L Cluster Label Candidates do
31: create cluster C described with L;
32: add to C all documents whose similarity to C exceeds the Snippet Assignment Theshold;
33: end for
34: put all unassigned documents in the "Others" group;

{STEP 5: Final Cluster Formation}

35: for all clusters do
36: clusterScore labelScore × C;
37: end for

ARCHITECTURE:

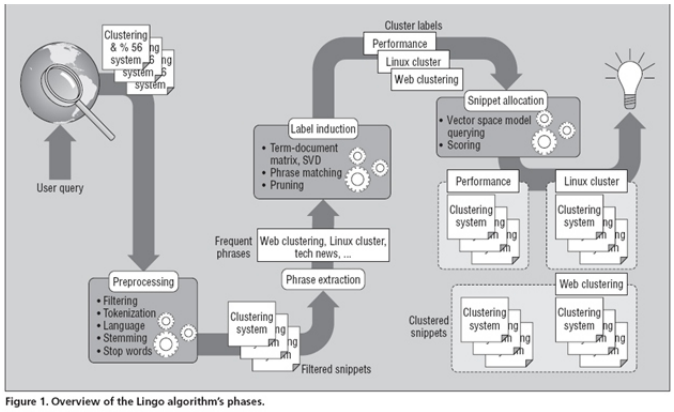

Figure 1. Overview of the Lingo algorithm's phases.

PHASES OF THE LINGO ALGORITHMS:

1. Preprocessing:

At this stage, we typically use a combination of three common text preprocessing methods-

- **Stemming:**

a technique for finding asemantic representation of an inflected word (usually a lemma) to decrease the impact of a language's syntax;

- **Ignoring stop words:**

a common technique for dealing with terms that occur frequently but have no meaning (conjunctions, articles, and so on); and

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

- **Text-segmentation:**

heuristics, a techniquefor dividing text into words and sentencesthat has many implementations.Phrase extractionthe phrase-extraction phase aims to discoverphrases and single terms that couldpotentially explain the verbal meaningsbehind the SVD-derived abstract concepts.Like the online semantic hierarchical clustering(SHOC) algorithm.

2.Cluster-label induction:

During the cluster-label-induction phase, Lingo identifies the abstract concepts that best describe the input snippet collection and uses frequent phrases to construct a human-readable representation of these concepts. This produces a set of labels, each of which will determine one cluster's content and description.

3. Cluster-content allocation:

The cluster-content allocation process resembles VSM-based document retrieval except that instead of one query, Lingo matches the input snippets against a series of queries, each of which is a single cluster label.

4. Test data and the experiment:

We took our ground truth and test data from the Open Directory Project (http://dmoz.org),a human-collected directory of Web page links and descriptions. Documents and groups (clusters) inside ODP are a result of the commonsense agreement of many people and one individual's subjective choice. In addition, unlike most classic information retrieval test suites, which contain full documents, ODP contains only short descriptions of documents, which serve as snippet replacements.

5. Empirical evaluation:

We manually investigated each cluster's contents and label for every test set at the 0.250 threshold level. Table 3 presents the topmost labels. Cluster descriptions were generally satisfactory ("federated data warehouse" and "foot orthotics," for example), even if elliptical because of truncated sentences in the input snippets ("information on infrared [photography]").

6. Analytical evaluation:

We can numerically compare similarity between two cluster structures in several ways—for example, using mutual-information measures.6 But these measures usually. Attempt to aggregate similarity between individual clusters into a single figure, whereas we wanted to show the differences in allocation of objects to clusters between Lingo and the suffix tree clustering (STC) algorithm.

The algorithm simulates a user navigating randomly in the Web who jumps to a random page with probability q or follows a random hyperlink (on the current page) with probability 1-q. It is further assumed that the user never goes back to the previously visited page following an already traversed hyperlink backwards. This process can be modelled with a Markov chain, from where the stationary probability of being in each page can be computed. This value is then used as part of the ranking mechanism.

CONCLUSION:

Each algorithm has its own merits and demerits. Lingo produces high cluster diversity; the Small outliers are highlighted well.In Lingo the number of clusters produced is more when compared to other two algorithms.We have presented a novel algorithm for clustering of Web search results. The inspiration for the algorithm was taken from both existing scientific work, and a commercial system. Our algorithm, however, took a different path in many areas. Specially, our contribution is in presenting a description of algorithm to our best knowledge.

Lingo achieves impressing empirical results, but the work on the algorithm is obviously not varnished Cluster label phase could be improved by adding elements of linguistic reorganization of nonsensical phrases. Topic separation phase currently requires computationally expensive algebraic transformations incremental with small memory footprint would be of great importance for algorithm. Finally, a more elaborate evaluation technique will be necessary to establish weak points in the algorithm.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

REFERENCES:

1. Oren Zamir and Oren EtzioniDocument Clustering: Feasibility Demonstration Proceedings of the 19th International ACM SIGIRConference on Research and Development of Information Retrieval,1998, pp 46-54.

2. Oren Zamir and Oren EtzioniGrouper: A Dynamic Clustering Interface to Web Search Results. WWW8/ Computer Networks,Amsterdam, Netherlands, 1999.

3. Oren E. Zamir. Clustering Web Documents: A Phrase-Based Method for Grouping Search Engine Results. Doctoral Dissertation, Universityof Washington, 1999.

4. Scatter/gather a cluster based approach to browsing large document collections Douglassr cutting,David R.Karger ,Jan O Pederson,15 annual International SIGIR 92,ACM 0-89791-542-0912/0006/0318.

5. Antonio Di Marco and RobertoNavigli, Clustering Web Search Results with Maximum Spanning Trees other publication details.

6. Ke,W., Sugimoto, C.R., Mostafa, J.: Dynamicity vs. effectiveness: studying online clustering for scatter/gather. In: Proc. of SIGIR 2009, MA, USA, 2009, pp. 19–26 7.Carpineto, C., Osinski, S., Romano.

7. Kamvar, M., Baluja, S.: A large scale study of wireless search behavior: Google mobile search. In: Proc. of CHI 2006, New York, NY, USA, 2006, pp. 701–709.

8. Osinski, S., Weiss, D.: A concept-driven algorithm for clustering search results. IEEE Intelligent Systems 20(3), 2005, 48–54.

9. ClusteringWeb-Search Results Using Transduction-Based RelevanceModelLurong Xiao and Edward Hung Department of Computing, The Hong Kong Polytechnic University, Hong Kong fcslxiao,csehungg@comp.polyu.edu.hk.

10. Sanderson, M.: Ambiguous queries: test collections need more sense. In: Proc. of SIGIR 2008, Singapore, 2008, pp. 499–506.

11. Chen, J., Zane, O.R., Goebel, R.: An unsupervised approach to cluster web search results based on word sense communities. In: Proc. ofWI-IAT 2008, Sydney, Australia, (2008),. pp. 725–729.

12. Zhang, X., Hu, X., Zhou, X.: A comparative evaluation of different link types on enhancing document clustering. In: Proc. of SIGIR 2008, Singapore, 2008,. pp. 555–562.

16. Incremental document clustering for webpage classification,Wai-Chiu Wong and Ada Wai-CheeFu,Dept of Computer science and Engineering,The Chinese University Hong Kong July 1,2000.

17. A New algorithm for clustering search results Giansalvatore Mecca,SalvartoreRaunich Alessandro Pappalardo Department of Mathematics and Informatics ,university of Basilicata,Potenza,ItalyApril 3,2007[6]

18. Clustering and Diversifying Web Search Results with Graph-Based Word Sense Induction Antonio Di Marco_ Sapienza University ofRome Roberto NaviglIDipartimento di Informatica, SapienzaUniversità di Roma, Via Salaria, 113, 00198 Roma Italy.