International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 2 Issue 6, June - 2013

# Inclusion Of Synonyms Concept In Lingo Algorithm For Improving Web Search Results Clustering

Pragna Makwana
Prof. Neha Soni

*Department of Computer Engineering, SVIT Vasad, Gujarat, India*

## *Abstract*

*The primary goal of web search engines is to help the Internet users to locate resources of interest on the Web. Web search results clustering is an increasingly popular technique for providing useful grouping of web search results, or snippets, into clusters. The Lingo algorithm, proposed by Stanislaw Osi´nski and Dawid Weiss, uses frequent phrases to identify candidate cluster labels, and then assigns snippets to these labels. This paper extends lingo algorithm by adding synonyms concept for tokens found in the lingo algorithm. Synonyms concept will help in two ways.*

1. *The documents having similar meaning but different words are included in cluster, which are excluded before synonyms concept added.*

2. *Overlapping clusters removed. Means words and their synonyms will not create different clusters.*



## 1. Introduction

With the increased use of internet we have large amount of shared information on World Wide Web. To access small piece of relevant information from this largest repository is overwhelming. Even with the use of search engines, it is difficult to find the most relevant documents from the returned list of large number of documents in response to the user query.

Clustering seems to be a promising approach to making the search results more understandable to the users. *Search results clustering* attempts to solve this problem by identifying and labeling groups of similar search results, and presenting this grouped output to the user as clusters.

Search results clustering is becoming increasingly popular; examples include commercial systems such as Vivisimo and IBoogie and research frameworks such as Carrot2 (http://project.carrot2.org/index.html).

Various search results clustering algorithms are available.AHC is simple but not very robust towards outliers and are slow when applied to large collection of documents. K-Means is very efficient and simple but very sensitive to input parameters.STC uses phrases to provide concise and meaningful description of groups. SHOC uses LSI and phrases in the process of clustering but it provides vague comments on the values of thresholds of the algorithm and the method which is used to label the resulting clusters.[2] In this approaches documents were assigned to cluster based on some mathematical properties, but problem is we know that certain documents should be clustered together but the relation between them cannot be explained. In Lingo this problem can be resolved by description comes first approach. But still the quality of the clusters created by lingo can be improved by adding synonyms concept in it. To add synonyms concept we used word net database, which is a lexical database for English dictionary, for finding synonyms of the tokens found in the lingo.

## 2. Lingo Algorithm

[1][3]The general idea behind Lingo is to first find meaningful descriptions of clusters and then based on the descriptions, determine their content. Lingo contains 5 phases as following:

### Phase 1: Pre-processing:

This phase includes common operations such as Filtering, Language identification, Tokenization stemming and stop word marking that improve the quality of the input snippets, and therefore of the frequent phrase detection and cluster labeling.

In the proposed work changes done in pre - processing step of lingo. Synonyms concept is added in preprocessing step of lingo algorithm as shown in the figure. For all tokens identified by tokenization, synonyms are found from wordnet database. Wordnet database is the lexical database for English language. In

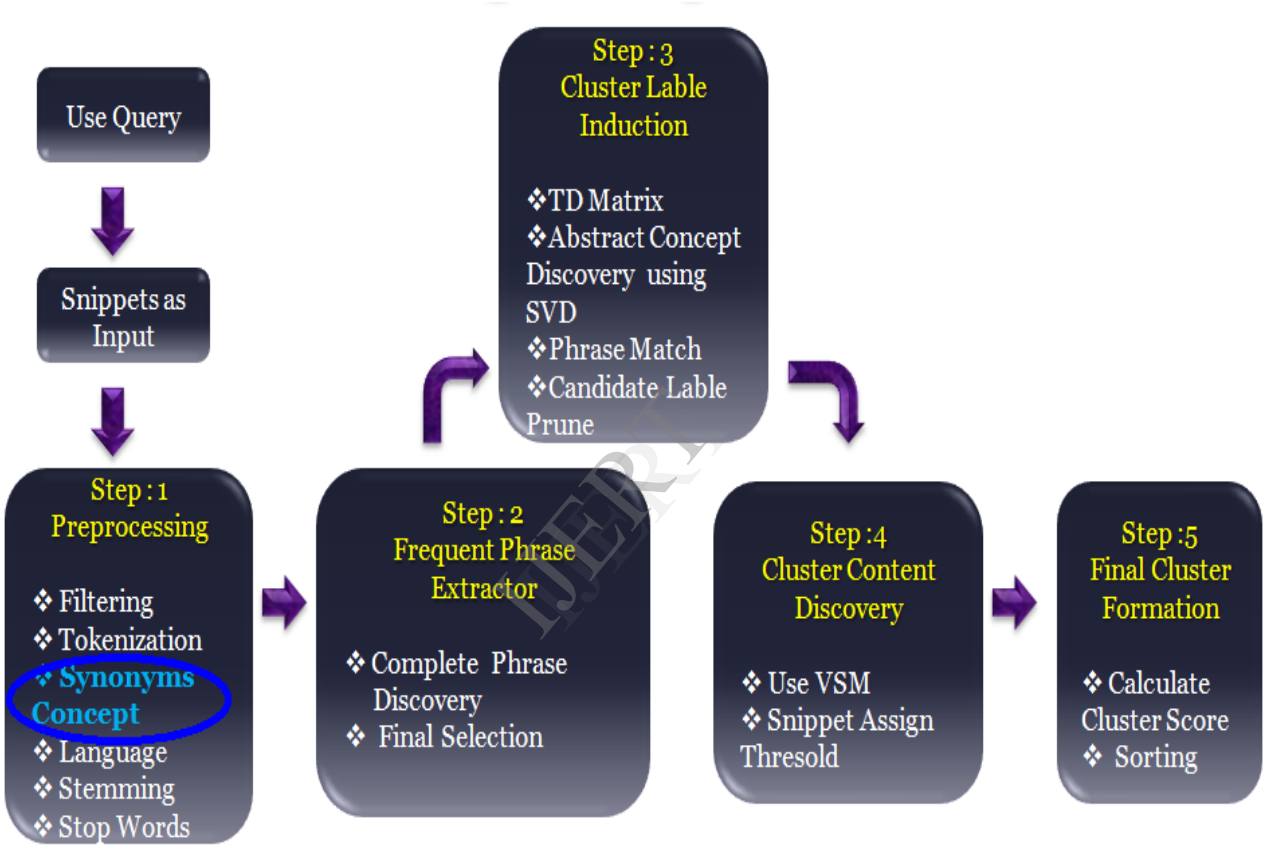


**Figure 1:** Flow of Proposed Lingo Algorithm

which nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Words in each synsets are same in their conceptual meaning [8]. For example the conceptual meaning in wordnet is "an evil supernatural being" and the words which are related by this meaning are "devil", "demon". Also "demon" is word for conceptual meaning "someone extremely diligent or skilful" in wordnet.

The proposed work adds synonyms from the word net and then it finds the most frequent synonym of that word and represents the word with that synonym.

### Phase 2: Frequent Phrase Extractor:

In this phase frequently occurring words and phrases are found according to certain criteria one of them is term frequency threshold. It is the minimum no of counts for term in the document. If the term frequency

satisfies the term frequency threshold then it can be included in frequently occurring terms.

### Phase 3: Cluster Label Induction:

In this phase for frequently occurring terms term document matrix is built and weight is calculated by tf-idf formula. Then Singular Value decomposition is applied to find abstract concepts from the documents. SVD is a method for identifying and ordering the dimensions along which data points exhibit the most variation. Once we have identified where the most variation is, it's possible to find the best approximation of the original data points using fewer dimensions. Hence, SVD can be seen as a method for data reduction[7]. After Abstract concept discovery phrase matching is done. It finds the best matching phrase which can represent the abstract concept. After candidate label pruning is performed. And overlapping clusters are removed.

### Phase 4: Cluster Content Discovery:

In the cluster content discovery phase, the classic Vector Space Model is used to assign the input snippets to the cluster labels induced in the previous phase. The assignment process much resembles document retrieval based on the VSM model – the only difference is that instead of one query, the input snippets are matched against every single cluster label. Snippet assignment to the cluster is depends on the Snippet Assignment Threshold.

### Phase 5: Final Cluster Formation:

In the final phase of the algorithm, cluster scores are calculated according to the following formula: Cluster-score = label-score * member-count. If in the presentation interface the resulting clusters are sorted according to their score, the user will presented with the well-described and relatively large groups in the first place.

## 3. Experimental Evaluation

Experiments are carried out on the following input data. Input snippet data is prepared from www.google.com and Open Directory Project (http://www.dmoz.org/), which is a human-categorized directory of internet resources, presented as URLs, document titles and snippets describing the site content.

**Table 1:** Document snippets used for experimental evaluation

| | **INPUT** |
|---|---|
| **0** | "http://www.islamawareness.net/Angels/",<br>"Angels in Islam",<br>"Islam mandates belief in Angels.<br>There are many kinds of Angels and<br>Quran and Hadith describes them in detail." |
| **1** | "http://whereangelswalk.com",<br>"Do You Believe in Angels?",<br>"Angels from a Biblical perspective<br>including stories, poems and links<br>to buy angel related products." |
| **2** | "http://inspirationalstories.com/angels-1.html",<br>"Afterhours Inspirational Stories",<br>"Inspiring stories about angels<br>and their presence." |
| **3** | "http://saints.sqpn.com/saintc25.htm",<br>"Patron Saints Index: Cuthbert",<br>"Illustrated details the saint." |
| **4** | "http://saints.sqpn.com/saint-christopher",<br>"Christopher",<br>"Profile of the popular saint,<br>takes legend as fact. Illustrated." |
| **5** | "http://www.deliriumsrealm.com/97/asmodeus",<br>"Asmodeus",<br>"From a Demons. Information<br>on the demon Asmodeus." |
| **6** | "http://www.qlddemons.com/",<br>"Queensland Demons",<br>"Official Website of teh<br>Queensland Demons." |
| **7** | "http://futurecam.com/dustDevils.html",<br>"Dust Devils",<br>"Short introduction to dust devils,<br>with a picture." |

Before adding synonym concept,original lingo has created 5 clusters and created separate clusters for "angel" and "saint" though they are synonyms in word net. Also it has not included 7$^{th}$ document in "demons" cluster though "devil" is synonym of "demon".

**Table 2:** Before Adding Synonyms Concept

**Output: 5 Clusters Created**

| |
|---|
| **Cluster 1: Angels (3 Docs)** |
| [0] Angels In Islam<br>[1] Do You Believe In Angels?<br>[2] Afterhours Inspirational Stories |
| **Cluster 2: Demons (2 Docs)** |
| [5] Asmodeus<br>[6] Queensland Demons |
| **Cluster 3: Saint (2 Docs)** |
| [3] Patron Saints Index: Cuthbert<br>[4] Christopher |
| **Cluster 4: Stories (2 Docs)** |
| [1] Do You Believe In Angels?<br>[2] Afterhours Inspirational Stories |
| **Cluster 5: Other (1 Doc)** |
| [7] Dust Devils |

After adding synonym concept, lingo has created 2 clusters and created one common cluster for "angel" and "saint" as they are synonyms. Also it has included 7th document in "demons" cluster as "devil" is synonym of "demon".

**Table 3:** After Synonyms Concept Added

**Output: 2 Clusters Created**

| |
|---|
| **Cluster 1: Angels (5 Docs)** |
| [0] Angels In Islam<br>[1] Do You Believe In Angels?<br>[2] Afterhours Inspirational Stories<br>[3] Patron Saints Index: Cuthbert<br>[4] Christopher |
| **Cluster 2: Demons (3 Docs)** |
| [5] Asmodeus<br>[6] Queensland Demons<br>[7] Dust Devils |

## 4. Conclusions and Future Work

As observed in the previous section, the inclusion of synonym concept using word synonyms from word net database has provided a clear improvement over the original lingo algorithm. The clear improvement is shown mainly in the assignment of documents to the clusters part and number of clusters. Here the overlapping clusters are removed. And documents having similar meaning are included in same cluster.

Still further enhancements can be carried out in lingo algorithm. Important enhancement to lingo will include, creating a hierarchical structure of clusters, either directly inferred from the source, or using man-made ontology such as Word Net.

## 6. References

[1] Stanislaw Osi´nski and Dawid Weiss," *A Concept-Driven Algorithm for Clustering Search Results*" , Pozna´n University of Technology2005,IEEE.

[2] K. Sridevi,R. Umarani, V.Selvi, "*An Analysis of Web Document Clustering Algorithms*", Department of Computer ScienceS,Nehru Memorial College,IJST,2011.

[3] Stanisław Osiński, "*An Algorithm For Clustering Of Web Search Results*",Poznań University of Technology, Poland, 2003.

[4] Stanisław Osiński, "*Dimensionality Reduction Techniques For Search Results Clustering*", Department of Computer Science The University of Sheffield, UK,2004.

[5] Stanislaw Osi´nski and Dawid Weiss, "*Conceptual Clustering Using Lingo Algorithm: Evaluation on Open Directory Project Data*", Institute of Computing Science, Pozna´n University of Technology.

[6] Stanis law Osi´nski, Jerzy Stefanowski, and Dawid Weiss, "*Lingo: Search Results Clustering Algorithm Based on Singular Value Decomposition*", Institute of Computing Science, Pozna´n University of Technology,Poland.

[7] Kirk Baker, "*Singular Value Decomposition Tutorial*", March 29, 2005

[8] *WordNet - Princeton University Cognitive Science Laboratory. 02 Jan. 2009 <http://wordnet.princeton.edu/>.*