International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 1 Issue 10, December- 2012

# Contextual Web Search Results Clustering Using Lingo And Extended Wordnet

Rathi Kiran
Alpha College Of Engineering and Technology,
Ahmedabad,India.

Mitula Pandya
Alpha College Of Engineering and Technology,
Ahmedabad,India.

Suryakant Vishwakarma
Birla Institute Of Technology & Science,
Pilani,India.

## Abstract

*Organizing Web search results into clusters ease users' quick browsing through search results. Traditional clustering techniques are least adequate as they don't suggest clusters with highly readable names. In the proposed paper we revaluate the clusters labels suggested by contextual recognition algorithm of lingo on the basis of eXtended WordNet database. Revaluation of cluster depends on the QTFactor(Quantity Factor) returned by the lookup for synonyms snippets in eXtended WordNet database. During the process of revaluation the eXtended WordNet database involves rebuilding using STANDUP lexical database API for lookup and extJWNL (Extended Java WordNet Library) Java API for updating the WordNet dictionary. With the invlolvement of above said APIs a better result clustering can be achieved.*

***Keyword***: *DataMining, Cluster Analysis, Feature extraction, Indexing.*

## 1. Introduction

Nowadays, many people are using internet it is the main source of information, it is suffering from overload. Web search engines like Google, bing and Yahoo can be considered as a cornerstone service for any Internet user. The keyword-based, boolean search style used by these engines has rapidly permeated user habits, to such an extent that it is now extending to other classes of applications, for example desktop search. The real problem arises when information expressed by search query is vague, too broad, or ill-defined, in which a huge list of irrelevant documents are returned back by the search engine. The ranking algorithm established by search engine involves a vital role in return result. Our approach is reverse of traditional order of cluster content discovery. Search result clustering try to solve this sort of problem by looking and labelling groups of simpat(similar-pattern) search and finally presenting the grouped output as a cluster to the end user.

Nowadays, Search results clustering is gaining a popularity in commercial systems such as Vivisimo (http://www.vivisimo.com), and IBoogie (http://iboogie.tv/), and research frameworks such as Carrot2 (http://project.carrot2.org/index.html).

The proposed paper extends the original Lingo algorithm proposed by Stanislaw Osi´nski and Dawid Weiss, and uses the approach of identifying frequent phrases as candidate cluster labels, then assigning snippets to these labels. This paper contributes in adding contextual recognition to enable the recognition of synonyms in snippets, thus improving the quality of the clusters generated.The contextual recognition is achieved using the eXtended WordNet database, which is a lexical database for the English language and advancement over WordNet database. eXtended WordNet holds "Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept".

## 2. Related Work

There are several commercial search engines that indulge some form of clustering. Besides Vivisimo and Grokker, other examples are Ask.com , iBoogie, Kartoo and WiseNut.Vivismo is based on document clustering and metasearch software automatically categorizes search results into hierarchical clusters.XML and XSL standards are used in Vivismo. Grokker, a visual search tool that clusters

related search results together in conceptually related categories.iBoogie is a search site which combines metasearch and clustering to deliever and organize result into structure content. The concept of clustering search results as a means to improve retrieval performance has been investigated quite deeply in information retrieval. A different clustering algorithms have been proposed that use learning to improve the cluster label generation process. Regarding contextual search results clustering, which is the title of this paper, a similar approach to what is provided in this paper is discussed in, which uses synonyms and hypernyms from eXtended WordNet in order to improve text document clustering.

## 3. Basic Concepts: Information Rerieval Techniques

### 3.1 Vector Space Model(VSM)

Vector Space Model (VSM) is a technique of information retrieval that transforms the problem of comparing textual data into a problem of comparing algebraic vectors in a multidimensional space. Once the transformation is done, linear algebra operations are used to calculate similarities among the original documents. Each component of the vector reflects a particular key word or term connected with the given document. The value of each component depends on the degree of relationship between its associated term and the respective document. Term weighting is a process of calculating the degree of relationship (or association) between a term and a document. Let aij represent the degree of relationship between term i and document j.

Term frequency weighting

$\mathbf{a_{ij} == tf_{ij}}$

where denotes how many times term i occurs in document j
Tf-idf weighting (term frequency inverse document frequency)

$\mathbf{a_{ij} == tf_{ij}. Log(N/df_i)}$

where $tf_{ij.}$ The term frequency $df_i$ denotes the number of documents in which term i appears and N represents the total number of documents in the collection.

In VSM, a collection of d documents described by t terms can be represented as a t x d matrix A, which is referred to as the term-document matrix.

$$\begin{array}{c} \\ T_1 \\ \vdots \\ T_m \end{array} \begin{array}{c} \begin{array}{ccc} D_1 & \cdots & D_n \end{array} \\ \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} \end{array}$$

Each element in the matrix aij represents the degree of relationship between term i and document j .

### 3.2 Latent Semantic Indexing

Latent Semantic Indexing is a technique of feature extraction which attempts to reduce the rank of a term-frequency matrix in order to get rid of noisy or synonymous words. An algebraic method of matrix decomposition called Singular Value Decomposition is used for discovering the orthogonal basis of the original term-document matrix. Performance data shows that these statistically derived vectors are more robust indicators of meaning than individual terms This basis consists of orthogonal vectors that, at least hypothetically, correspond to topics present in the original term-document matrix.SVD breaks a t × d matrix A into three matrices U, _ and V , such that A = U_V T . U is a t × t orthogonal matrix whose column vectors are called the left singular vectors of A, V is a d × d orthogonal matrix whose column vectors are called the right singular vectors of A, and _ is a t × d diagonal matrix having the singular values of A ordered decreasingly along its diagonal. The rank rA of matrix A is equal to the number of its non-zero singular values. The first rA columns of U form an orthogonal basis for the column space of A—an essential fact used by Lingo.These frequent words are considered as cluster labels.

## 4. Clustering Algorithm: Lingo Algorithm and its Modifications.

The Lingo algorithm combines common phrase discovery and latent semantic indexing techniques to separate search results into meaningful groups. It looks for meaningful phrases to use as cluster labels and then assigns documents to the labels to form groups.

In the Lingo description-comes-first approach, careful selection of label candidates is crucial. The algorithm must ensure that labels are significantly different while covering most of the topics in the input snippets. To find such candidates, we use the vector space model (VSM) and singular value

decomposi-tion (SVD), the latter being the fundamental mathematical construct underlying the latent semantic indexing (LSI) technique

## 4.1 Lingo Algorithm Phases

### 4.1.1 Pre-processing:

This phase include operation such as Text Filtering, Language Identification, Stemming, Stop word Marking that improves the quality of snippets which results good phrase detection and cluster labelling.

### 4.1.2 Feature Phrase Extraction

The aim of the feature extraction phase is to discover phrases and single terms that will potentially be capable of explaining the verbal meaning behind the LSI-found abstract concepts.
To be considered as a candidate for a cluster label, a phrase or term must:
• Appear in the input documents at least a specified number of times.
• Not cross sentence boundaries.
• Be a complete phrase Compared to partial phrases.
• Not begin nor end with a stop word.

### 4.1.3 Cluster label Induction

Once frequent phrases (and single frequent terms) that exceed term frequency thresholds are known, they are used for cluster label induction. In the cluster label induction phase, meaningful group descriptions are formed based on the SVD decomposition of the term-document matrix. There are four steps to this phase: term-document matrix building, abstract concept discovery, phrase matching and label pruning and evaluation.

### 4.1.4 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.

### 4.1.5 Final cluster formation

Finally, clusters are sorted for display based on their score, calculated using the following simple formula:

Cscore = label score × kCk

## 4.2 Contextual clustering using eXtended WordNet

The Lingo algorithm was carried out in the frequent phrase extraction phase described above. The modification carried out in this paper adds an extra step that involves finding the synonyms of the frequent terms and phrases.

### 4.2.1 Two plugable API has been introduced to increase the Quality of Cluster Label

1) **STANDUP lexical database API** provides access to the lexical database functionality of the STANDUP system and allows you to fetch the synonyms of the word from a vast database which keeps on updating.

2) **extJWNL(Extended Java WordNet Library):** extJWNL (Extended Java WordNet Library) is a Java API for creating, reading and updating dictionaries in WordNet format. extJWNL is an upgraded version of JWNL.

**Algorithm 1.** Pseudo-code of modified Extract Single Term

```
1: D <- input documents (or snippets)
2: T <- single terms
3: F <- list of Features, empty
4: TD <- list of Term-Document Arrays, empty
5: for each document d in D
6: for each frequent term t in T
7: if t does not exist in F then
8: td <- new Term-Document Array
9: increase term frequency in td for document d by
   1
10: add td to TD
11: f <-new Feature
12: increase total term frequency for term t by 1
13: add t, td to f
14: add f to F
15: else // t exists in F
16: td<- Term-Document Array for term t
17: increase term frequency in td for document d by1
18: f <- Feature that contains term t
19: increase total term frequency for term t by 1
20: end if
21: end for
22: end for
23: // end of original Extract Single Terms
24: // start adding synonyms
25: SF<- list of synonym Features, empty
26: STD <- list of synonym Term-Document Arrays,
empty
27: for each document d in D
28: for each term t in F
29: find word for term in eXtended WordNet
30: if word is found
```

```
31: for each synset syn for word
32: for each synonym in syn
33: if synonym is the same as t // avoid
    adding the original word from the synset
34: continue to next synonym
35: end if
36: if synonym is found in F
37: td <-Term-Document Array for synonym
38: increase term frequency in td for
   document d by 1
39: f <- Feature that contains term synonym
40: increase total term frequency for term
    synonym by 1
41: else // synonym not found in F
42: if synonym is found in SF // check
for same condition above in previous synonyms
43: std<- Term-Document Array for Synonym
44: increase term frequency in std for
    document d by 1
45: sf <- Feature that contains term synonym
46: increase total term frequency for term
    synonym by 1
47: else // synonym not found in F or SF
48: td <- new Term-Document Array
49: increase term frequency in td for
    document d by 1
50: add td to STD
51: f <-new Feature
52: increase total term frequency for
    term synonym by 1
53: add synonym, td to f
54: add f to SF
55: end if
56: end if
57: end for
58: end for
59: end if
60: end for
61: end for
62: append SF to F
63: append STD to TD
```

# 5. Experimental Evaluation

## 5.1 Input Data Preparation

We selected document from the ODP for our experiment .Open Directory Project (ODP) is a tree-like, human-collected thematic directory of resources in the Internet. Each branch of this tree, called a category, represents a topic and contains links to resources in the Internet that relate to this topic. Short description is added to each topic which serves as a substitute for snippets. We selected random categories for our experiment. We decided to add document that have synonyms in order for the test of Contextual Lingo Algorithm as it is based on eXtended WordNet. The clustering results were compared with the original Lingo algorithm and Contextual Lingo algorithm.

**Table 1. Document topics used for experimental evaluation**

| Document Topics |
|---|
| Theory Of Architecture. |
| Designers are Innovators. |
| Painters bring your imagination to life. |
| Artist works from soul. |
| Introduction to modern data retrieval. |
| Professional Architect. |
| Designers are new Rockstar. |
| Software for sparse singular value decomposition |
| Matrix Computation |
| Creature driven by daemons. |

## 5.2 Results

Comparison of the clusters generated using the original Lingo algorithm to the clusters generated after the modification to inclusion of synonyms in the frequent phrase extraction phase.

**Table 2. Comparison of Orignal Lingo Algorithm and Conceptual Lingo Algorithm**

| Original Lingo Algorithm | Contextual Lingo Algorithm |
|---|---|
| Cluster-1: Designer<br>1.Designer are Innovators<br>2.Designers are new Rockstar | Cluster-1:Architecture<br>1.Theory of Architecture<br>2.Designer are Innovators<br>3.Professional Architecture |
| Cluster-2: Painter<br><br>1.Painters bring your imagination to life. | Cluster 2:Artist<br>1.Creature driven by daemons<br>2.Painters bring your imagination to life<br>3.Artists work from soul |

The differences can be realized if it is noted that the following groups of words are synonyms in eXtended WordNet:

1. “Designer” and “Architecture”
2. “Painter” and “Artist”

Contextual Lingo recognize the document “Designers are Innovators” because designer and architecture are synonyms in eXtended WordNet. The same concept was applicable in cluster-2 “Painter brings your imagination to life” because painter and artists are

synonyms. So by this experiment evaluation we came to know that original Lingo Algorithm did not included concept of synonyms.

## 6. Conclusion and Future Work

With the inclusion of contextual recognition using synonyms from eXtended WordNet leverage has been provided over the original Lingo algorithm, broadly in the assignment of documents to clusters where the matching was improved because of the knowledge of relations between words. This has been further enhanced with the plugging of 2 APIs (STANDUP lexical database API & extJWNL (Extended Java WordNet Library)).However, some future experimentation and analysis needs to be carried out to assess the cluster results before the cluster is rebuilt.

1. **Enclosing Quality Factor and Quantiy Factor:**
Revaluation of the cluster depends on the QTFactor, QTFactor stands for Quantity as well as Quality.In the proposed paper we have included the Quantity Factor as the QTFactor. Algorithms can be developed so that Quality of clusters can be identified and depending on it eXtended WordNet database can be revaluated.

**2. Allowing more lexical API's:**
With the inclusion of lexical API's the eXtended Wordnet library is expanded with more and more Quality words and better content. eXtended WordNet assuming that the original term is a noun. For the purposes of this experiment, this has been sufficient to evaluate the improvement of including eXtended WordNet synonyms however a complete approach would include identification of the Part of Speech of the original term.

## 7. Reference

[1] Stanisław Osiński “An Algorithm for Clustering of websearch Results” Master Thesis poznań university of technology, poland, june 2003

[2]Oikonomakou, Nora, and Michalis Vazirgiannis. "A Review of Web Document Clustering Approaches." Data Mining and Knowledge Discovery Handbook.

[3] Stanislaw Osinski and Dawid Weiss “A Concept Driven Algorithm for Clustering Search Results” IEEE Intelligent Systems, 2005

[4] Poonam C. Fafat, Prof.S.S.Sikchi “Lingo an approach for Clustering” International Journal of Engineering Reseach & Technology Vol.1 Issue3,May- 2012

[4] Ahmed Sameh, Amar Kadray “Semantic Web Search Results Clustering Using Lingo and WordNet” International Journal of Reseach and Reviews in Computer Science Vol.1 No.2 ,June 2010

[5] K.sridevi,R.Umarani,V.Selvi”An Analysis of Web Document Clustering Algorithms” international Journal of Science and Technology vol.1 No.6 ,December 2011

[6] Stanis law Osi´nski, Jerzy Stefanowski, and Dawid Weiss“ Lingo: Search Results Clustering Algorithm Based on Singular Value Decomposition” Poznań University of Technology, Poland .

[7] J. Stefanowski and D. Weiss, “Carrot2 and Language Properties in Web Search Results Clustering,” Proc. Web Intelligence, 1st Int’lAtlantic Web Intelligence Conf. (AWIC 2003),LNCS 2663, E.M. Ruiz, J. Segovia, and P.S.Szczepaniak, eds., Springer, 2003.

[8] Osinski, Stanislaw. "Improving Quality of Search Results Clustering with Approximate Matrix Factorizations." 28th European Conference on IR Research (ECIR 2006), 10 Apr. 2006,London, UK. Springer Lecture Notes in Computer Science. Vol. 3936. 2006. 167-78.

[9] Osinski, Stanislaw, and Dawid Weiss. "Conceptual Clustering Using Lingo Algorithm: Evaluation on Open Directory Project Data." Advances in Soft Computing, Intelligent Information

[10] eXtendedWordNet http://xwn.hlt.utdallas.edu/