Contextual Search Results Clustering Using Lingo with Synonymity

Engineering

KEYWORDS : Data mining, Clustering, Search Results Clustering, Stemming

Rathi Kiran PG Student, Computer Engineering Department, Alpha College Of Engineering , Khatraj, Ahmedabad, India.

Mitula Pandya Assistant Professor, Computer Engineering Department, Alpha College Of Engineering , Khatraj, Ahmedabad, India.

ABSTRACT

The Web search results returned as snippets are arranged into clusters ease user's quick browsing. Classical clustering techniques are least adequate as they don't suggest clusters with highly readable names. In the proposed paper we revaluate the document assigned to cluster suggested by contextual recognition algorithm of lingo on the basis of synonmity of cluster depends on the qtfactor(quantity factor). With the involvement of synonymity a better result clustering can be achieved.

1. INTRODUCTION

The internet contains vast information and majority of publicly available search engines but a linear ranking of documents need to be cluster. Search results clustering is a process of organizing document references returned by a search engine into a number of meaningful thematic categories.

Search results clustering have several advantages over the classical text clustering. Firstly search results clustering is based on short document excerpts returned by the search engine called snippets. Secondly, as the clustering algorithm is intended to be a part of an on-line search engine, the thematic groups must be created ad hoc and fully automatically. Finally, as the main objective of search results clustering is to help the users to identify the documents of interest more quickly.

We here represent Lingo as Search results clustering algorithm which emphasis on the good quality of cluster label. The proposed paper extends the Lingo Algorithm by the adding contextual recognition of synonyms in snippets, thus improving quality of cluster generated. The modification in Lingo Algorithm are done using synonymity with the inclusion of Extended Wordnet which have increased semantic relations between the documents and also quality of labels.

2. Related Work

There are various algorithm used for search result clustering differing from each other. This section introduces the related studies by describing some of them.

The [1] describes Lingo algorithm combines common phrase discovery and latent semantic indexing techniques to separate search results into meaningful groups. The [2] proposes a linear weighted method of Single-Pass improvement, which integrates HowNet semantic similarity and cosine similarity, fuses and rediscovers clusters, and extracting the cluster labels. The [3] enhances search results by performing fuzzy clustering on web documents returned by conventional search engines, as well as ranking the results and labeling the resulting clusters. This is done using a fuzzy transduction-based clustering algorithm (FTCA).

The [4] enhanced Lingo algorithm by Suffix Array Similarity Clustering (SASC) for clustering web search results. This method creates the clusters by adopting improved suffix array, which ignores the redundant suffixes, and computing document similarity based on the title and short document snippets returned by Web search engines.

3. LINGO: SEARCH RESULTS CLUSTERING ALGORITHM

In the Lingo description-comes-first approach, is described. The algorithm must ensure that labels are significantly different while covering most of the topics in the input snippets. The phases of Lingo described as below.

3.1 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 labeling.

3.2 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.

3.3 Cluster Label Induction

Once frequent phrases 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.

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

3.5 Final Cluster Formation

Finally, clusters are sorted for display based on their score.

4. Methodology

This section describes the enhanced methodology. Lingo algorithm is enhanced in pre-processing step by inclusion of synonyms. The primary aim of the preprocessing phase is to remove from the input documents all characters and terms that can possibly affect the quality of group descriptions.

4.1 Text filtering

In the text filtering step, all terms that are useless or would introduce noise in cluster labels are removed from the input documents..

4.2 Tokenization

Tokenization is a process of identifying word and sentence boundaries in a text. The simple tokenizer could us white space character7 as word delimeter and selected puntuation marks as '.' As sentence boundaries.

4.3 Language identification

Before proceeding with stemming and stop words marking, for each input document separately, LINGO tries to recognize its language. In this way, for each snippet, appropriate stemming algorithm and stoplist can be selected.

4.4 Stemming

In this step, inflection suffixes and prefixes are removed from each term appearing in the input collection. This guarantees that all inflected forms of a term are treated as one single term, which increases their descriptive power. We use a free Java implementation of Porter stemmer for English. For example the words: connected, connecting, interconnection are treated as a single word connect.

5 Stop words marking

The words which have no meaning in the document such as and, the are removed.

The pseudo code to modify pre-processing with synonym is described. For each document do the text filtering, perform tokenization, identify language then stemming, but apply stemming with synonym. The case normalization of a word is found then calculate frequency statistics for words. In addition to that also calculate term frequency and term frequency by document that is used in stemming and also in further phases. While performing stemming, the task is to perform stemming of a word but here if we find synonym of a word then perform stemming of a synonym else of a word.

Table 1: Pseudo Code to Modify pre-processing

Pseudo Code to Modify Preprocessing with Synonym

```
For Each Document
{
1. Do Text Filtering;
2. Performs Tokenization for Documents;
3. Apply Document's Language;
4. Apply Stemming With Synonym;
 4.1 Performs Case Normalization And Calculates Frequency Statistics for Words;
      4.1.1 Calculate Term Frequency (tf) of the Word;
      4.1.2 Calulate Term Frequency By Document (tfByDocument);
//Applies Stemming to Word;
 4.2 Find Synonym Of Word;
 If Synonym exists
       Perform Stemming Synonym for Word;
else
       Perform Stemming for Word;
 4.3 Convert Lists To Arrays And Store them in Stems;
      4.3.1 Sum of Term Frequency (tf) of the Word and Synonym.
      4.3.2 Concate Term Frequency By Document (tfByDocument) of Word and Synonym.
5. Mark Stop Words;
}
```

5. EXPERIMENTATION RESULTS

In this section, we include the input used for our implementation. Input is the user query used for search the documents and then we allocate snippets returned by search results as a documents to algorithm which are further processed on enhanced Lingo algorithm. Firstly enlisted user defined statements that are taken in previous papers and applied enhanced algorithms and compared the results. Further implemented algorithm on 100 documents, that are search results taken from [15]. Carrot2 is a freely available Open Source framework for experiments with processing and visualization of search results which works on Lingo algorithms and is known as clustering engine. We have downloaded the source of lingo algorithm from github [16].

Table 2: Documents

No.	Document
1	Large Scale Singular Value Computations
2	Software for the Sparse Singular Value Decomposition
3	Introduction to Modern Information Retrieval
4	Linear Algebra for Intelligent Information Retrieval
5	Matrix Computations
6	Singular Value Analysis of Cryptograms
7	Data Retrieval Organization

The document that have synonyms are added in order for the test of Contextual Lingo Algorithm as it is based on Extended WordNet. The synonym file includes synonyms such as data & information. The clustering results were compared with the original Lingo algorithm and enhanced Lingo algorithm.

In the comparison , the document named Data Retrieval organizations was added to the cluster Information Retrieval with Enhanced Lingo because of semantic relation between documents. So, number of documents assigned to cluster was increased and other topics results decreased.

Table-3: Clusters using Lingo

no.	Created 4 clusters using Lingo
1	Singular Value (3 docs, score: 2.55)
	[0] Large Scale Singular Value Computations
	[1] Software for the Sparse Singular Value Decomposition
	[5] Singular Value Analysis of Cryptograms
2	Computations (2 docs, score: 2.29)
	[0] Large Scale Singular Value Computations
	[4] Matrix Computations
3	Information Retrieval (2 docs, score: 5.04)
	[2] Introduction to Modern Information Retrieval
	[3] Linear Algebra for Intelligent Information Retrieval
4	Other Topics (1 docs, score: 0)
	[6] Data Retrieval Organization

This section shows the result of clusters using Enhanced Lingo in which 3 clusters are created and their score is described.

Table-4: Clusters using Enhanced Lingo

no.	Created 3 clusters using Enhanced Lingo
1	Information Retrieval (3 docs, score: 2.97)
	[2] Introduction to Modern Information Retrieval
	[3] Linear Algebra for Intelligent Information Retrieval
	[6] Data Retrieval Organization
2	Singular Value (3 docs, score: 1.98)
	[0] Large Scale Singular Value Computations
	[1] Software for the Sparse Singular Value Decomposition
	[5] Singular Value Analysis of Cryptograms
3	Computations (2 docs, score: 2.32)
	[0] Large Scale Singular Value Computations
	[4] Matrix Computations

Below , the graph shows the comparison of Lingo and Extended Lingo, which shows that quality of cluster label of "Information Retrieval" has increased.

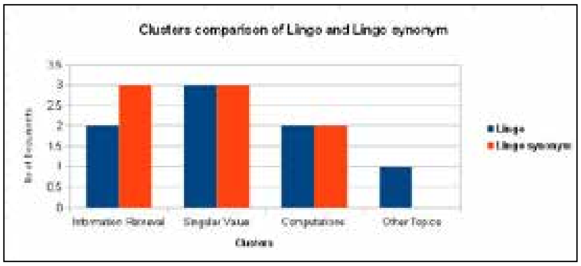

Figure 1: Cluster comparison of Lingo and Enhanced Lingo of 7 documents

Further, implemented the enhanced Lingo on 100 documents of data mining from carrot2. The graph is shown below, which shows a leverage in Enhanced Lingo Algorithm.

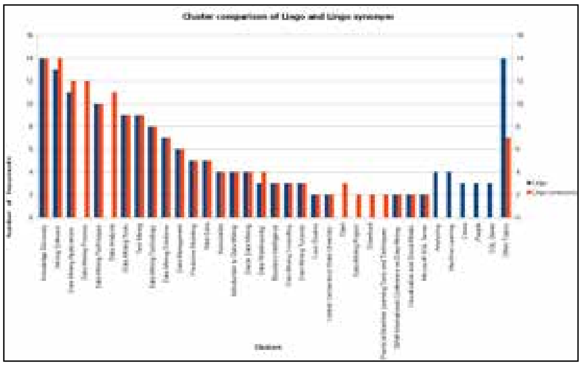

Figure 2: Cluster comparison of Lingo and Enhanced Lingo of 100 search Results

6. CONCLUSION AND FUTURE WORK

In the Lingo-Enhanced Algorithm, we have included synonyms in the Pre-Processing stage. With the inclusion of synonyms a leverage has been provided to 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. The Algorithm yields a better Qulalitized and Quantized result set. The default cluster size has been reduced.

The Algorithm can be further extended to include Near-synonyms. Near-synonyms preserve truth-conditions, or propositional meaning, to a level of granularity of representation consistent with language-independence in most contexts when interchanged. To update the Synonyms on the fly Extended-Word-net can be utilized.

REFERENCE

[1] Stanislaw Osinski and Dawid Weiss "A Concept Driven Algorithm for Clustering Search Results "IEEE ,June 2005 | [2] Dequan Zheng, Haibo Liu, TiejunZhao "Search Results Clustering Based on a Linear Weighting Method of Similarity",IEEE -2011 | [3] Takazumi Matsumoto, Edward hung "Fuzzy Clustering and Relevance Ranking of Web Search Results with Differentiating Cluster Label Generation "IEEE-2010 | [4] Shunlai Bai,Wenhao Zhu, Bofeng Zhang, Jianhua Ma "Search Results Clustering Based on Suffix Array and VSM" IEEE-2010. |[5] Michael Steinbach, George karypis, Vipin Kumar "A Comparison of Document Clustering Techniques". | [6] Rizwan Ahmad, Dr.Aasia Khanum "Document Topic Generation in Text Mining by Using Cluster Analysis with E- ROCK" International journal of Comput er Science and Security..(IJCSS) vol.4 iss 2. | [7] Rekha Baghel, Dr Renu Dhir "A Frequent Concepts Based Document Clustering Algorithm" International Journal Of Computer Applications, July 2010. | | [8] Somjit Arch-int"Web Document Clustering using Semantic Link Analysis" IEEE -2005. |[9] J.Han and M.Kimber,2000 .DataMining: Concepts and Techniques. | [10] Stanislaw Osinski, Jerzy Stefanowski and Dawid Weiss," Lingo: Search Results Clustering Algorithm Bases on Singular Value Decomposition", Institute of Computer Science, Poznam University of Technology,2004 | [11] R.Subhashini and .Jawahar Senthil Kumar, "The Anatomy of Web Search Clustering and Search Engines", Indian journal of Computer Science and Engineering Vol 1,No.4 | [12] Sedding,Julian and Dimitar kazakov." WordNet –based Text Document Clustering" |[13] B.C.M Fung ,K.Wan, M.Ester 2003, "Hierarchical Document Clustering using Frequent Itemsets",SDM'03. | | [14] Andreas Hotho, Andreas Nurnberger "A Brief Survey of Text Mining"May 13 ,2005 | [15] (http://search.carrot2.org/stable/search) | [16] https://github.com/lex-lingo/lingo |