Similarity Measurement Based Web Document Clustering for Query

Kishor Wagh
Department of Information Technology, Government College of Engineering, Amravati, India

Abstract

Document clustering is automatic organization of documents into clusters so that documents within a cluster have high similarity in comparison to documents in other clusters. It is measuring similarity between documents and grouping similar documents together. The study of similarity measure for clustering is initially motivated by a research on automated text categorization. There were different similarity measures. It provides efficient representation and visualization of the documents; thus helps in easy navigation also. It has been used intensively because of its wide applicability in various areas such as web mining, search engines, and information retrieval. The aim of organizing data in such a way is to improve data availability and to fasten data access, so that web information retrieval and content delivery on the web are improved. Tf-Idf based Apriori algorithm is carried out on search engine's result. The F-measure for the entire clustering is 0.8426 for query 'boy lad'.

Keywords: *Search engine, Ranking, Information Retrieval, Text Mining.*

1. Introduction

The process of discovery of new information generates large volumes of data that can be overwhelming if not properly stored and/or utilized. With the recent explosive growth of the amount of content on the Internet, it has become increasingly difficult for users to find and utilize information. Other large document repositories are growing so rapidly that it is difficult and costly to categorize every document manually. Problems can arise with classifications and interpretations because of subjective nature due to human judgments and different levels of training. It is important to provide a framework that utilizes text mining techniques in developing system. Web mining can apply intelligent methods to extract or mine knowledge and meaningful data patterns from a large amount of unstructured texts or documents for decision-making. In order to deal with these problems, research toward automated methods of working with web documents initiated so that they can be more easily browsed, organized, and indexed with minimal human intervention.

The application of document clustering [5] [1] [9] to information retrieval has been motivated by the potential effectiveness gains postulated by the cluster hypothesis. The hypothesis states that relevant documents tend to be highly similar to each other, and therefore tend to appear in the same clusters. Initially, document clustering was investigated for improving the precision or recall in information retrieval systems and as an efficient way of finding the nearest neighbors of a document. Document clustering is automatic organization of documents into clusters so that documents within a cluster have high similarity in comparison to documents in other clusters. It has been studied intensively because of its wide applicability in various areas such as web mining, search engines, and information retrieval. It is measuring similarity between documents and grouping similar documents together. It provides efficient representation and visualization of the documents; thus helps in easy navigation also.

2. Clustering System Architecture

System architecture [10] gives brief idea about what actual system is. System architecture is the conceptual model that defines the structure, behavior and more views of a system. Clustering system architecture is shown in Figure 1.

An architecture description is a formal description and representation of a system, organized in a way that supports reasoning about the structures of the system. System architecture comprises system components, the externally visible properties of those components, the relationships between them.

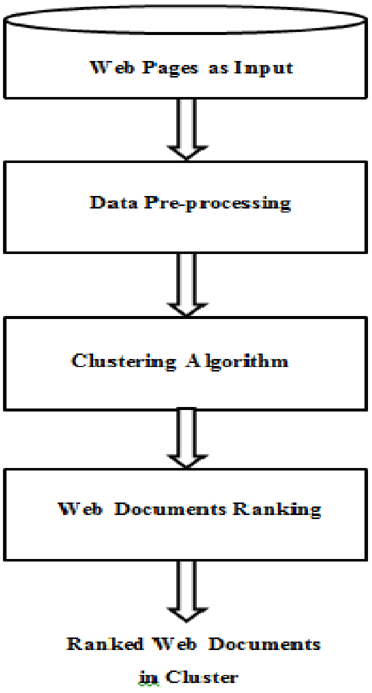

Figure 1: Architecture of Clustering System.

2.1 Web Pages as Data Input

Advances in data collection and storage capabilities during the past decades have led to an information overload. Researchers working in domains as diverse as engineering, astronomy, biology, remote sensing, economics, and consumer transactions, face larger and larger observations and simulations on a daily basis.
Much of the information presented to Internet users today is in HTML (Hyper Text Markup Language) tables. Indeed, tables are a useful way to organize information and display it effectively and attractively. It is needed to import HTML information. According to the system architecture web pages are input to the data mining model. As to provide input to the data mining model, web resources are used such as online web pages. A Search Engine Results Page (SERP) is the list of results that a search engine returns in response to a specific word or phrase query. Each listing includes the linked Web page title, the linked page URL (Uniform Resource Locator), a brief description of the page content and, in some cases, links to points of interest within the website. These online web pages are saved in the system for further processing. The number of online web pages are accessed the more system will improve. As a result system will require data storage capacity as large as possible so that as many as web pages can be save. Collection of more web pages may take more time for processing.

2.2 Data Pre-processing

Web pages consists of massive volume of data which will make data tends to be inconsistent and noisy. If data is inconsistent, then there is possibility that mining process can lead to confusion which results in inaccurate results. In order to extract data which is consistent and accurate data pre-processing is applied on that data. Preprocessing text is called text normalization. The objective of this is that it enhances the quality of data and at the same time reduces the difficulty of mining process.
The pre-processing task on web pages is much more complex than any traditional text document collection. Only a small portion of the Webs pages contain truly relevant or useful information, dismissing the rest as uninteresting data that serves only to swamp the desired results.
Remove Stop Words: Stop word removal is common preprocessing step. The most common words are in any web page document does not provide meaning of the documents; those are prepositions, articles, and pro-nouns etc. These words are treated as stop words. Because every web page document deals with these

words which are not necessary for text mining applications. These words are eliminated. Any group of words can be chosen as the stop words for a given purpose. This process also reduces the text data and improves the system performance. Example: the, in, a, an, with etc.

Stemming: Stemming is a technique for the reduction of words into their root. The stem of a word is the base part. Many words in the English language can be reduced to their base form or stem e.g. agreed, agreeing, disagree, agreement and disagreement belong to agree. Stemming will lower the dimensionality of document by term matrix by turning cat and cats into the same term. The porter stemmeris the classic stemming algorithm. Furthermore are names transformed into the stem by removing the s. The variation Peters in a sentence is reduced to Peter during the stemming process. Words like look can be inflected with a morphological suffix to produce looks, looking, looked. They share the same stem look. It is beneficial to map all inflected forms into the stem. This is a complex process, since there can be many exceptional cases. The stem is useful and important, because all other inflections of the root are transformed into the same stem.

Preprocessing contains remove stop words and stemming. It gives preprocessed documents as output. The objective of this is that it enhances the quality of data and at the same time reduces the difficulty of mining process.

Feature Extraction: A document is usually represented as a document vector in which each component indicates the value of the corresponding feature in the document. The feature value can be term frequency, relative term frequency, or a combination of term frequency and inverse document frequency.

Similarity Measurement: It is a symmetric measure, the difference between presence and absence of a feature is considered more essential than the difference between the values associated with a present feature. The similarity increases as the difference between the two values associated with a present feature decreases. The similarity decreases when the number of presence-absence features increases.

2.3 Clustering Algorithm

The goal of clustering is to reduce the large amount of raw data by categorizing in smaller sets of similar items. Web page clustering puts together web pages in groups based on similarity or other relationship measures. Many algorithms are used for documents clustering such as K-means algorithm [5] [6] [3], Tf-Idf based Apriori.

Algorithm 1 Basic K-means Algorithm for finding K clusters.

```
1. Select K points as the initial centroids.
2. Assign all points to the closest centroid.
3. Recompute the centroid of each cluster.
4. Repeat steps 2 and 3 until the centroids don't change.
```

In this paper, introduced a mechanism called Tf-Idf based Apriori [5] [6] [8] for clustering the web documents. Tf-Idf based Apriori algorithm is carried out on search engine’s result. Rank the documents [7] in each cluster using Tf-Idf and similarity factor of documents based on the user query. This approach will help the user to get all his relevant documents in one place and can restrict his search to some top documents of his choice.

Cluster Formation Using Tf-Idf based Apriori:

Input: Minimum support and web documents
Output: Number of clusters of web documents

The steps of the algorithm are given as follows:

1. Web page extraction and preprocessing: Submit the query to a search engine and extract top ‘N’ pages. Preprocess the retrieve corpus as follows:
   - Remove the stop and unwanted words.
   - Select noun as the keywords from the corpus and ignore other categories, such as verbs, adjectives, adverbs and pronounce.
   - Do stemming using porter algorithm.
   - Save each preprocessed ‘N’ pages as documents D_i, where i = 1, 2, 3,. . . ,N .

2. After keyword extractions, consider each keyword as a transaction and the documents D_i in which the keyword occurs as transaction elements.

3. Calculate tf for each distinct keyword in each D_i as,

$$\text{tf}=\frac{\text{No. of times the keyword appears in the document}}{\text{Total number of keywords in the document}} \qquad (1)$$

   Calculate idf for each distinct keyword in each D_i as

$$\text{idf}=\log_{10}\frac{\text{total number of documents}}{\text{number of documents the keyword appears in}} \qquad (2)$$

4. Calculate tf* idf value for each distinct keyword in each D_i and represent all the values in the tf*idf table.

5. Calculate threshold as,

$$\text{threshold}=\frac{1}{\text{minimum support}}\times\log_{10}\frac{\text{total number of documents}}{\text{minimum support}} \qquad (3)$$

6. Generate n frequent candidate itemsets (S, where n>=2) for keywords till, 0 < min(tf*idfD_1 , tf*idfD_2 , …, tf*idfD_N) <= threshold for all generated S and at each step do the followings:

   Calculate tf as, tf = 1/ (number of times S appears in the document) for each n frequent candidate itemset in each document.

   Calculate idf as, idf = log10 (total number of documents/number of documents S appears in) for each n frequent candidate itemset.

   Calculate tf*idf value for each n frequent candidate itemset in each document and represent all the values in the tf*idf table.

Now mark the 'n' frequent candidate itemsets (rows) for elimination if min{ (tf *idfD$_1$, tf* idfD$_2$,…,tf*idfD$_N$)}> threshold.

Mark documents (columns) for elimination if min{(tf*idf n frequent candidate item set$_1$, tf*idf n frequent candidate item set$_2$, ... ,tf*idf n frequent candidate item set$_N$)}> threshold.

7. Final Clusters (C$_i$) where i = 1, 2, 3,. . . ,M formed each having group of similar documents(D$_k$) where k=1. . . N and N may vary from cluster to cluster.

Here, the clusters are formed when system uses the query 'boy lad' and the clusters have to be ranked.

Cluster 1:

1. Document 5:

   naar: a boy, lad, youth, retainer. Original Word Part of Speech: Noun Masculine Transliteration: naar. Phonetic Spelling: (nah'-ar) Short Definition: men. http://biblehub.com/hebrew/5288.html
2. Document 6:

   Lightning Lad was there alongside Cosmic Boy and Saturn Girl to ward off the Persuader and help stop Brainiac by any means necessary, including killing ...

   http://www.comicvine.com/lightning-lad/4005-1253/

3. Document 8:

   Nov 11, 2013 ... Can "lad" only be used to address a male, while "mate" both and female? ... Lad is another name for a boy or a young man. I suspect that ...
   http://english.stackexchange.com/questions/137091/whats-the-difference-between-lad-and-mate-in-british-english

4. Document 11:

   11 Items ... Discover the designer Good Lad Boys selection at Belk. Explore the name brand Good Lad Boys collection today and get free shipping deals. http://www.belk.com/products/good-lad-boys-Cb33855.jsp
5. Document 13:
   Welcome to Frisch's Big Boy! Store Locations Nearest You ... Super Big Boy. Super Big Boy. 1/2 lb. of beef* ... Brawny Lad . Brawny Lad . 1/4 lb. of beef* ... http://www.frischs.com/menu/menu.aspx

6. Document 14:
   Good Lad Baby Boys' 2-Piece Gingham Shirt Shorts Set. Orig. $ 32.00. Was $ 26.99 ... Good Lad Baby Boys' 3-Piece Sweater, Shirt Khakis Set. Orig. $ 58.00 http://www1.macys.com/shop/kids-clothes/baby-boy-clothes/Brand/Good% 2 0Lad?id=48693

7. Document 17:
   When I was a lad I served a term. As office boy to an Attorney's firm. I cleaned the windows and I swept the floor, And I polished up the handle of the big front

...
http://www.victorianweb.org/mt/gilbert/porter.html

Cluster 2:

1. Document 1:

   Information about lad in the free online English dictionary and encyclopedia.
   ... (ld). n. 1. A boy or young man. 2. Informal A man of any age; a fellow. http://www.thefreedictionary.com/lad
2. Document 4:

   a boy or young man. : a man with whom you are friendly. From Septon to Smallclothes: The real history behind 10 Game of Thrones words ...
   http://www.merriam-webster.com/dictionary/lad
3. Document 7:

   Lightning Lad is a founding member of the Legion of Super-Heroes along with Saturn Girl and Cosmic Boy. Born on the planet Winath, he is the twin brother of ... http://en.wikipedia.org/wiki/Garth

4. Document 10:

   ... matter, Hagar? Do not be afraid; God has heard the boy crying as he lies there. ... you, Hagar? Do not fear, for God has heard the voice of the lad where he is.
   http://biblehub.com/genesis/21-17.htm

5. Document 12:

   Lard Lad Donuts is a donut shop in the town of Spring eld, The store's mascot,
   ... The name and the statue of the eponymous boy are likely references to Big Boy ...
   http://simpsons.wikia.com/wiki/LardLadDonuts

6. Document 15:

   De ne lad and get synonyms. What is lad? lad meaning, pronunciation and more by Macmillan Dictionary. ... Men and boys: male, boy, young man. http://www.macmillandictionary.com/us/dictionary/american/lad

7. Document 16:

   Lyrics. Jack the lad. Lawrence in the desert. How was he to know? Under so much pressure from the men back home. Play with re you must be mad. Are you ...
   http://www.petshopboys.co.uk/lyrics/36/J

8. Document 19:

   There are 460 calories in 1 burger of Frisch's Big Boy Brawny Lad Burger. You'd need to walk 120 minutes to burn 460 calories. Visit CalorieKing to see calorie ...
   http://www.calorieking.com/foods/calories-in-sandwiches-burgers-brawny-lad-b urger-f-ZmlkPTIwOTk1Nw.html

9. Document 20:

a boy or young man; (informal) a familiar form of address for any male; a lively or dashing man or youth (esp in the phrase a bit of a lad); a young man whose
...
http://www.collinsdictionary.com/dictionary/english/lad

Cluster 3:

1. Document 2:

   Informal. a familiar or a ectionate term of address for a man; chap. 3. British Horseracing Informal. a stable boy. Origin of lad. Expand. Middle English. late Old ...

   http://dictionary.reference.com/browse/lad

2. Document 9:

   Matter-Eater Lad is the fteenth member inducted into the Legion of Super-Heroes, joining soon after Bouncing Boy. In his rst appearance, Matter-Eater Lad ...

   http://en.wikipedia.org/wiki/Matter-EaterLad

2.4 Ranked Web Documents in Cluster

The last step is sorting of the documents according to their rank values. There are certain similar records that fall in one category or form one cluster. After sorting the documents in the cluster according to the ranked values, system returns the ranked documents in the clusters.

3. Results and Discussion

For the query `boy lad', three clusters are formed for the web documents. The first cluster contains the documents 5, 6, 8, 11, 13, 14, 17. The second cluster contains the documents 1, 4, 7, 10, 12, 15, 16, 19, 20. And the third cluster contains the documents 2 and 9.
Precision and recall values for clusters formed of documents (snippets for query word `boy lad') are calculated. The precision and recall values for clusters are shown in Figure 2.

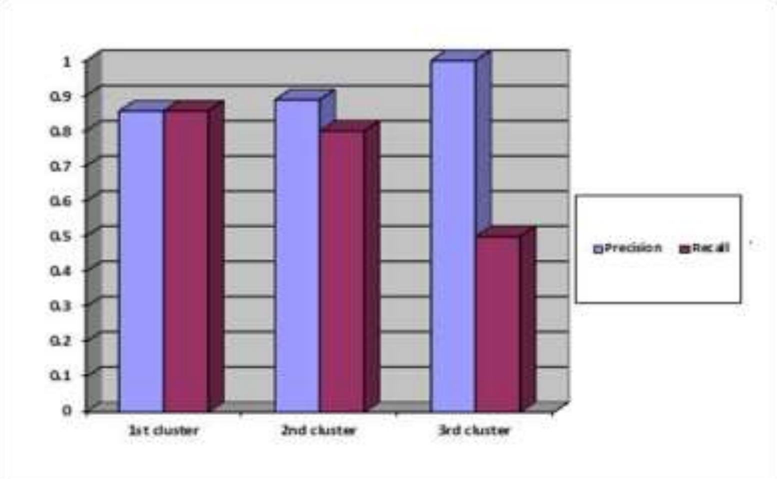

Figure 2: Precision and Recall Values for the Clusters of Dataset `boy lad'.

Here, System show the F-measures of the clusters formed after submitting a particular query. In this paper, four queries are executed - `boy lad' (Dataset 1), `pepsi' (Dataset 2), `microsoft' (Dataset 3), `ipod' (Dataset 4). The snippets that are received are treated as datasets for the clustering system. Figure 3 shows the F-measure values for different datasets for the queries. The F-measure for the entire clustering is 0.8426 for query ‘boy lad’.

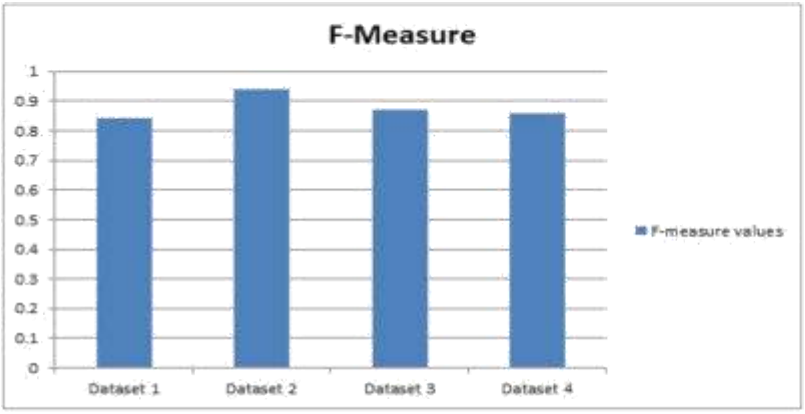

Figure .3: F-measures for Different Datasets.

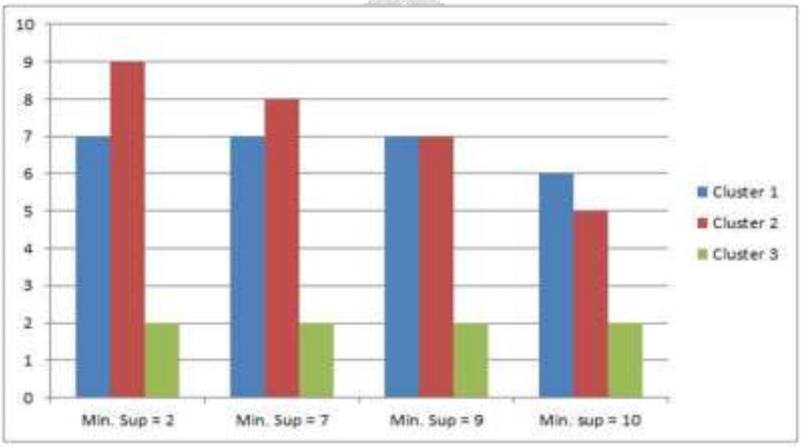

Figure 4: Cluster-Sets for Different Values of Minimum Support.

Now observe the clustering system by changing the basic parameter value, i.e. minimum support. We are calculating the threshold value on the basis of minimum support. Figure 4 shows the cluster sets for different values of minimum support for the dataset for query `boy lad'. It changes the size of cluster.

4. Conclusion

By optimizing similarity measures the optimal clusters can be formed thus performance is improved. Overall aim is organizing data in such a way that to improve data availability and to fasten data access, so that web information retrieval and content delivery on the web are improved.

The F-measure for the entire clustering is 0.8426 for query ‘boy lad’. Minimum support changes the size (number of documents) of cluster. We observed the clustering system by changing the basic parameter value, i.e. minimum support.

References

[1] Krishna Sapkota,Laxman Thapa,Shailesh Pandey Efficient Information Retrieval using measures of Semantic Similarity,2006.

[2] Anna Formica Concept similarity by evaluating information contents and feature vectors: A combined approach. Communications of the ACM, Vol.52, 2009.

[3] Leacock C.,Chodorow M.,”Combining local context and WordNet similarity for word sense identification”,In Fellbaum 1998,pp.133-138.

[4] George Tsatsaronis and Vicky Panagiotopoulou, A Generalized Vector Space Model for Text Retrieval Based on Semantic Relatedness,2009.

[5] Yung-Shen Lin, Jung-Yi Jiang and Shie-Jue Lee. A Similarity Measure for Text Classification and Clustering. IEEE Transactions On Knowledge And Data Engineering, 2013.

[6] Rajendra Kumar Roul, Omanwar Rohit Devanand, Sanjay Kumar Sahay. Web Document Clustering and Ranking using Tf-Idf based Apriori approach. IJCA Proceedings on International Conference on Advances in Computer Engineering and Applications ICACEA (2):34-39, March 2014.

[7] Kishor Wagh, Satish Kolhe, Improving Web Link Mining using Semantic Similarity Measurement, International Journal of Applied Engineering Research (IJAER), Volume 9, Number 19 (2014),2014, ISSN 0973-4562 , PP. 5663-5677.

[8] Kishor Wagh, Satish Kolhe, Evaluate Semantic Similarity of Words Using Semantic Distance, VNSGU Journal of Science and Technology Vol. 3, Issue 2, March 2012, ISSN: 0975-5446, PP. 22-29.

[9] Kishor Wagh, Satish Kolhe, Information Retrieval Based on Semantic Similarity Using Information Content, IJCSI International Journal of Computer Science Issues, Vol. 8 ,Issue 3,May 2011, ISSN(Online): 1694-0814,PP. 364-370 .

[10] Kishor Wagh, Satish Kolhe, Semantic Similarity Based on Information Content, International Journal of Computer Science and Application, Issue 2010, ISSN: 0974-0767, PP. 82-85.

JETIR