International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 4 Issue: 1 193 - 195

Design an Approach for Finding the Similarity between the Documents

Miss. Shilpa A. Satone
Prof. Jayant Rohankar

***Abstract*:** Information on internet is very huge in size. Web users need support to manage information easily. This makes the user's time consuming because there are many near-duplicate results. The efficient detection of near-duplicate articles is very important in many applications that have a large amount of data. We introduce algorithms of extracting key phrase and matching signatures for near-duplicate articles detection. Based on N-gram (i.e. bigram & trigram) algorithm for key phrase extraction & jaccard similarity for finding similarity between documents. Algorithms are applied on article. Results show that our proposed methods are more effective than other existing method.

***Keywords*:** *keyphrase, similarity, extraction, near-duplicate.*

*****

I. Introduction:

Information on internet is very huge in size. Web users need support to manage information easily. Search engines become the major breakthrough on the web for retrieving the information. Search engine will return closest results according to user's request. The web user has to go through the long list and inspect the titles, and snippets sequentially to recognize the required results. This makes the user's time consuming because there are many near-duplicate results. The efficient detection of near-duplicate articles is very important in many applications that have a large amount of data.

The main goal of this paper is to extract keyphrases and detect duplicate article in a particular field based on similarity using Bngram & jaccard similarity measure. Algorithms are applied on News Debate. Experimental results show that our proposed methods are effective.

The significant increase in number of the online newspapers has given web users a giant information source. The users are really difficult to manage content as well as check the correctness of articles.

II. Brief Literature Survey:

Paper [1]:- The 2015 IEEE RIVF International Conference on Computing & Communication Technologies Research, Innovation, and Vision for Future (RIVF)
Domain-Specific Keyphrase Extraction and Near-Duplicate Article Detection based on Ontology

This paper Based on ontology, keyphrases of articles are extracted automatically and similarity of two articles is calculated by using extracted keyphrases. Algorithms are applied on Vietnamese online newspapers for Labor & Employment. Experimental results show that our proposed methods.

Paper [2]:- Fei Xie1, 3, Xindong Wu1, 2 2, Proc. 9th IEEE Int. Conf. on Cognitive Informatics (ICCI'10) F. Sun, Y. Wang, J. Lu, B. Zhang, W. Kinsner & L.A. Zadeh (Eds.) 978-1-4244-8040-1/10/$26.00 ©2010 IEEE, Keyphrase Extraction Based on Semantic Relatedness.

This paper Based on the document is represented as a relatedness graph. Keyphrases are extracted based on the semantic relatedness features acquired from the graph. Our experiments demonstrate that the proposed keyphrase extraction method always outperforms the baseline methods TFIDF and KEA. Furthermore, our approach is not domain-specific and the method generalizes well when it is trained on one domain (journal articles) and tested on another (news web pages).

Paper [3]:- Junping Qiu, Qian Zeng, 978-1-4244-5824-0/$26.00 c 2010 IEEE, Detection and Optimized Disposal of Near-Duplicate Pages

This paper analyzed the existing algorithms to select an appropriate algorithm to detect near-duplicate pages, and optimized the disposing strategy to ensure that near duplicate pages would not take up too much space in search results while being used effectively. These will allow users to retrieve needed information more easily.
Keywords

III. Problem Formulation

The problem in existing system is that it gives less accurate result of precision & recall.

Objectives

The primary objectives of this study can be summarized as follows:

- To extract key phrase from document.
- Find similarities between the documents.
- To optimize the result so that it will gives a proper output.

IV. Research Methodology/Planning of Work:

In the proposed system we use N-gram(i.e. bigram & trigram) algorithm for key phrase extraction & jaccard similarity for finding similarity between documents.

The main goal of this approach is to find emerging topics in post streams by comparing the term frequencies from the current time slot with those of preceding time slots. We

propose the metric which introduces time to the classic score. This approach indexes all keywords from the posts of the collection. In addition to single keywords, the index also considers bigrams and trigrams. Once the index is created, the score is computed for each-gram of the current time slot based on its document frequency for this time slot and penalized by the logarithm of the average of its document frequencies in the previous
time slots .

$$df - idf_t = \frac{df_i + 1}{\log\left(\frac{\sum_{j=i}^{t} df_{i-j}}{t} + 1\right) + 1}.$$

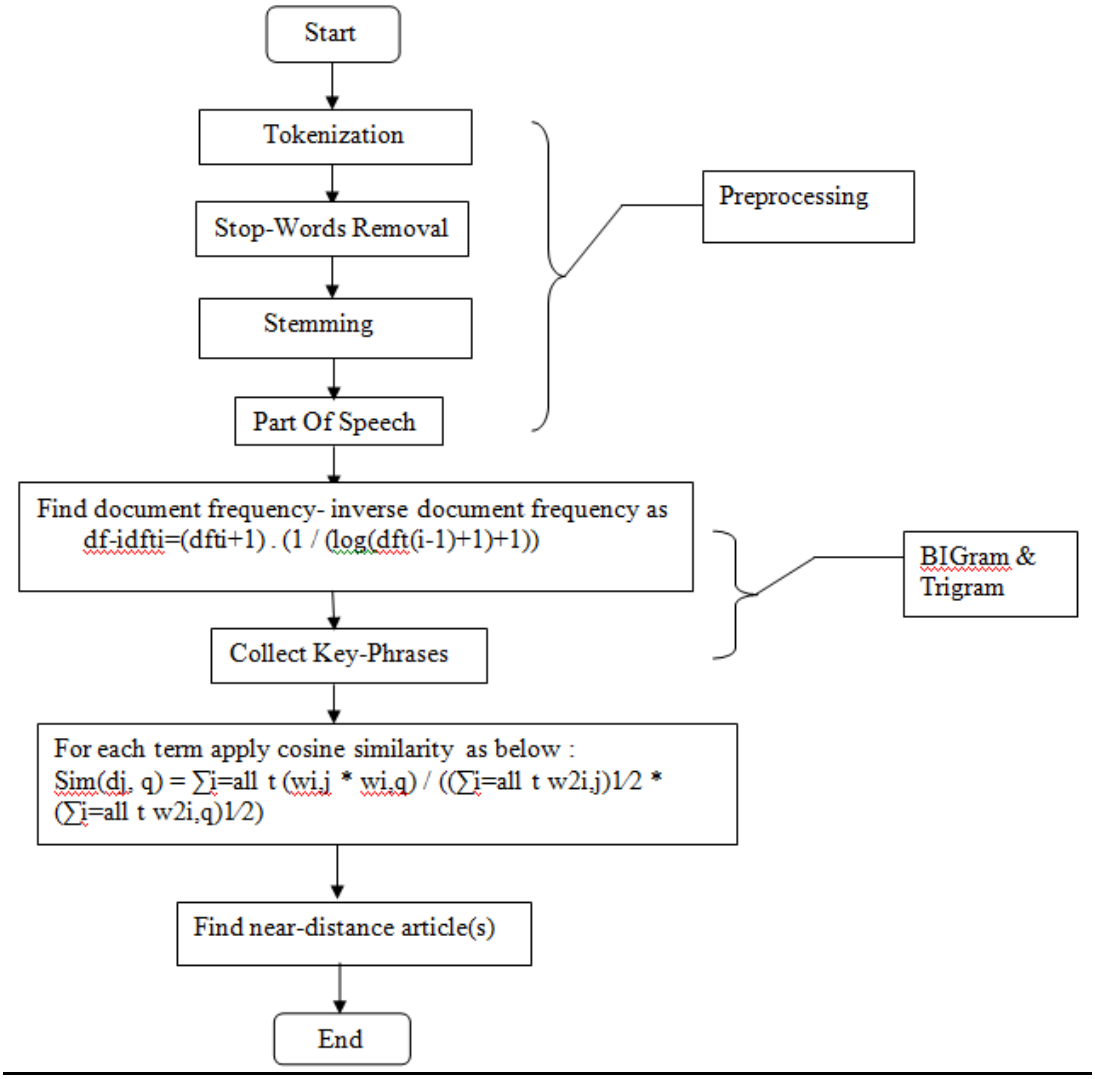
In the fields of computational linguistics and probability, an ***n*-gram** is a contiguous sequence of *n* items from a given sequence of text or speech. The items can be phonemes, syllables, letters, words or base pairs according to the application. The *n*-grams typically are collected from a text or speech corpus. When the items are words, *n*-grams may also be called **shingles**.

An *n*-gram of size 1 is referred to as a "unigram"; size 2 is a "bigram" (or, less commonly, a "trigram"); size 3 is a "trigram". Larger sizes are sometimes referred to by the value of *n*, e.g., "four-gram", "five-gram", and so on.

Trigrams do have an advantage over bigrams but it is small. Probably a similar order of magnitude decrease .Trigrams improve things 10% over bigrams.

Cosine Similarity Formula :
Sim (dj, q) = ∑i=all t (wi,j * wi,q) / ((∑i=all t w2i,j)½ * (∑i=all t w2i,q)½)
w i,j is weight that term i has for document j. dj is the document, represented here by the blue arrow. And since we have only one document, we can call j = 1. q is the query and is represented by the red arrow. t is the total number of terms in our space. In this very simple case t=3.

V. System Flow:



VI. Conclusion

Information on internet is very huge in size. Web users need support to manage information easily. This makes the user's time consuming because there are many near-duplicate results. The efficient detection of near-duplicate articles is very important in many applications that have a large amount of data. In above paper we learned different techniques and algorithms proposed by different authors over the years of research. From research we can see that this domain can still be improved for much better efficiency and accuracy of the system. In the proposed system we use N-gram(i.e. bigram & trigram) algorithm for key phrase extraction & jaccard similarity for finding similarity between documents. The main goal of this approach is to find emerging topics in post streams by comparing the term frequencies from the current time slot with those of preceding time slots. We propose the metric which introduces time to the classic score. This approach indexes all keywords from the posts of the collection. In addition to single keywords, the index also considers bigrams and trigrams.

VII. Bibliography:

[1] The 2015 IEEE RIVF International Conference on Computing & Communication Technologies Research, Innovation, and Vision for Future (RIVF) Domain-Specific Keyphrase Extraction and Near-Duplicate Article Detection based on Ontology

[2] Proc. 9th IEEE Int. Conf. on Cognitive Informatics (ICCI'10) F. Sun, Y. Wang, J. Lu, B. Zhang, W. Kinsner & L.A. Zadeh (Eds.) 978-1-4244-8040-1/10/$26.00 ©2010 IEEE, Keyphrase Extraction Based on Semantic Relatedness

[3] Junping Qiu, Qian Zeng, 978-1-4244-5824-0/$26.00 c 2010 IEEE, Detection and Optimized Disposal of Near-Duplicate Pages

[4] Bassma S Alsulami et al, International Journal of Computer Science & Communication Networks,Vol 2(2), 147-15, ISSN:2249-5789, Near Duplicate Document Detection Survey

[5] Wu, Y. et al, "Efficient near-duplicate detection for Q&A forum", in Proc. of 5th International Joint Conference on Natural Language Processing, pp. 1001-1009, 2011.

[6] Krishnamurthy Koduvayur Viswanathan and Tim Finin, "Text Based Similarity Metrics and Delta for Semantic Web Graphs", in Proceedings of the Poster Session of 9th International Semantic Web Conference, pp. 17-20, 2010.

[7] Shital Gaikwad, Nagaraju Bogiri "A Survey Analysis On Duplicate Detection in Hierarchical Data ", 2015 International Conference on Pervasive Computing (ICPC)