GSA: A GLOBAL FRAMEWORK FOR SIMILARITY SEARCHING

Anima Srivastava[1], Manish Jaiswal[2], Arpita Tewari[3]

[1]Department of Electronics & Communication, University of Allahabad, Allahabad, India
[2]Department of Electronics & Communication, University of Allahabad, Allahabad, India
[3]Department of Electronics & Communication, University of Allahabad, Allahabad, India

Abstract : With the advancement in technology searching and machine learning is believed to a good technique for measuring documents similarity and prediction accuracy for plagiarism detection. The most popular searching algorithm is either the industrial or the academic environment is RankBrain algorithm. This paper proposed an improved framework of searching with machine learning which masters the complexity of searching accurate matches. An empirical evaluation of the proposed approach is given based on its objective and case study. We describe a novel functional framework based on searching algorithm with machine learning both for differentiating intent of query and generate content semantically. We explore and analyze various well-known Google's searching algorithm in terms of their effectiveness toward similarity searching and best matching.

Index Terms - searching, document similarity, RankBrain

I. INTRODUCTION

Machine learning algorithms [16] are one of the powerful techniques to measure similarity of documents by versatile methods. This paper is carrying the actions of different similarity based machine learning algorithm but emphasizes on RankBrain i.e. the new way to design and find improved search ranking and quality. This work shows a transparent comparative study of similarity detection having its efficiency and deficiency in complete manner with analysis. The rest of the paper is organized as follows; Section 1 contains the introductory explanations of the work, Section 2 describes the brief knowledge of the several prominent contemporary searching algorithms, whereas the section 3 highlights the literature review of related searching [11] aspects and algorithms, section 4 stated clearly the detail of proposed framework; section 5 measuring the efficiency and applicability of the proposed framework; finally the section 6 includes the conclusion.

II. SEARCHING ALGORITHMS FOR SIMILARITY DETECTION

Each searching algorithm [6] has multiple parameters and searching criteria to detect similarity and retrieve optimum outcomes. Some of most popular Google's searching algorithms [7] are discussed in the following:

2.1Panda

Panda [7] is a searching algorithm used to assign grade for web pages which is based on subject's quality and also settle on down rank of websites with their quality content. Panda works like a strainer instead d of Google's other searching algorithm. Basically it is integrated into the ranking algorithm and used for de-rank sites with low quality content, it doesn't utilize in real time search but filtering and retrieving results from updated version of Panda is much more faster than the older one.

2.2Pigeon

Pigeon is Google's searching algorithm released with the two key factors i.e. distance and location. Pigeon is available for searches result in English only. The query is based on searcher's location because it significantly drops in the number of queries used to rank local and non local result returned. It uses local directory sites for providing excellent result. Goggle map and Google web search consistently used by pigeon for relevant local search results.

2.3 Penguin

The main objective of penguin [7] is to detect and de-rank sites with unsolicited, anomalous link outlines. By using devious tactics it operates in real time hence correction and revival takes less time penguin is just a segment of Google's main ranking algorithm.

2.4 Pirate

Google's pirate was invented inhibit and de-rank those sites that have many copyright encroachment reports. Nowadays popularly know sites are involved in making plagiarize content e.g. video clips, songs, movies etc. for attracting sightseers to download and surf freely. Now a day's various new sites appear with pirated content that would not be possible to rank them in a cluster that's why pirate algorithm separately down rank the websites with legal patent encroachment reports.

2.5 Fred

The name Fred coined by Google's 'Gary Illyes' suggested all updates done by Fred. Mostly blog pretentious sites with low quality content that often generate most majority of ads messages associated with direct or indirect revenue filtered by Fred In simple word, major function of Fred is to mark those unnatural content that generally disobey guidelines of Google web master.

2.6 Possum

The possum is used for local rank filtering. Possum enhances miscellaneous result i.e. based on physical location of the searcher. Furthermore, organizations that share an address with another organization of a similar type are degraded in the search result. Specially used for delivering better, more diverse results based on the searcher's location and the business' address

2.7 Mobilegeddon

Mobilegeddon is also referred to as Mobilepocalyse, Mopocalypse or Mobocalypse .Google uses Mobilegeddon to ensure that web pages optimized for mobile devices place at uppermost section and subsequently de-rank rest of pages mobile friendliness is a page-level factor of Mobilegeddon, means if one page of site can be considered as mobile friendly and ranked up while ranking are down ranked.

2.8 Hummingbird

Hummingbird [6, 9] search algorithm was invented to optimize the interpretation of searcher's query. The query may be of conversational talk or explained in longer way. Hummingbird focuses on the meaning behind the whole query rather than keywords uses within the query. The use if synonyms have also been considered by hummingbird instead of exact word match. Humming bird strengthen by semantic search capabilities or LSI [17] in place of syntactic significance.

2.9 PageRank

PageRank (PR)[6] gives each web page a rating according to user's interests and navigation time devotion on that web page. PageRank of any web page is actually the probability that a random surfer will go to the particular web page and how much time spend over there. Basically it depends on the current state of user's interest not on its history. PageRank effectively measures user's interest of surfing web pages and compute ranking of pages that is based on Markov process, where system moves like user from one state to other state depends on probability information in which likelihood of changing from one state to other possible state.

2.10RankBrain

RankBrain[6,14] is intellectual searching algorithm that helps Google to refine query processing and serve best matching result Google uses RankBrain to determine what result appear on the Google's search page and evaluate the relevancy of search result. Sometimes, RankBrain is called as machine-learning artificial intelligence system or it is a part of Google's overall search algorithm. According to "Bloomberg" RankBrain uses artificial intelligence to acquire knowledge both from being recognize the intent and from generating the content. Specially used to deliver better search results based on relevance & machine learning with artificial intelligence

III. LITERATURE REVIEW

D.Hassana [1] has detected semantic similarity by comparing and calculating two application using normalized compression distance.

M. Fowke et al [2] has discussed the simhash algorithm for the semantic similarity and for fulfilling this objective they use pinger software.

H. Wael [3] discussed the existing works on text similarity through partitioning them and concerned samples are presented.

I.FUJINO and Y. HOSHINO[4] has used Japanese morphological analysis to pickup terms from twitter data and calculate tf-idf to provide width parameter for each term and finally calculate document similarity from weight parameter vector between any two documents.

S.A.Hiremath and M.S.Otari [5] has described various plagiarism detection methods and compare them on the basis of their characteristics and performance; they have also use different plagiarism detection software to find it.

IV. The Proposed Approach

An efficient global search algorithm technique for RankBrain [14] has been developed to provide readability effectiveness and make the relevancy in better way. The framework of the suggested methodology is as follows:

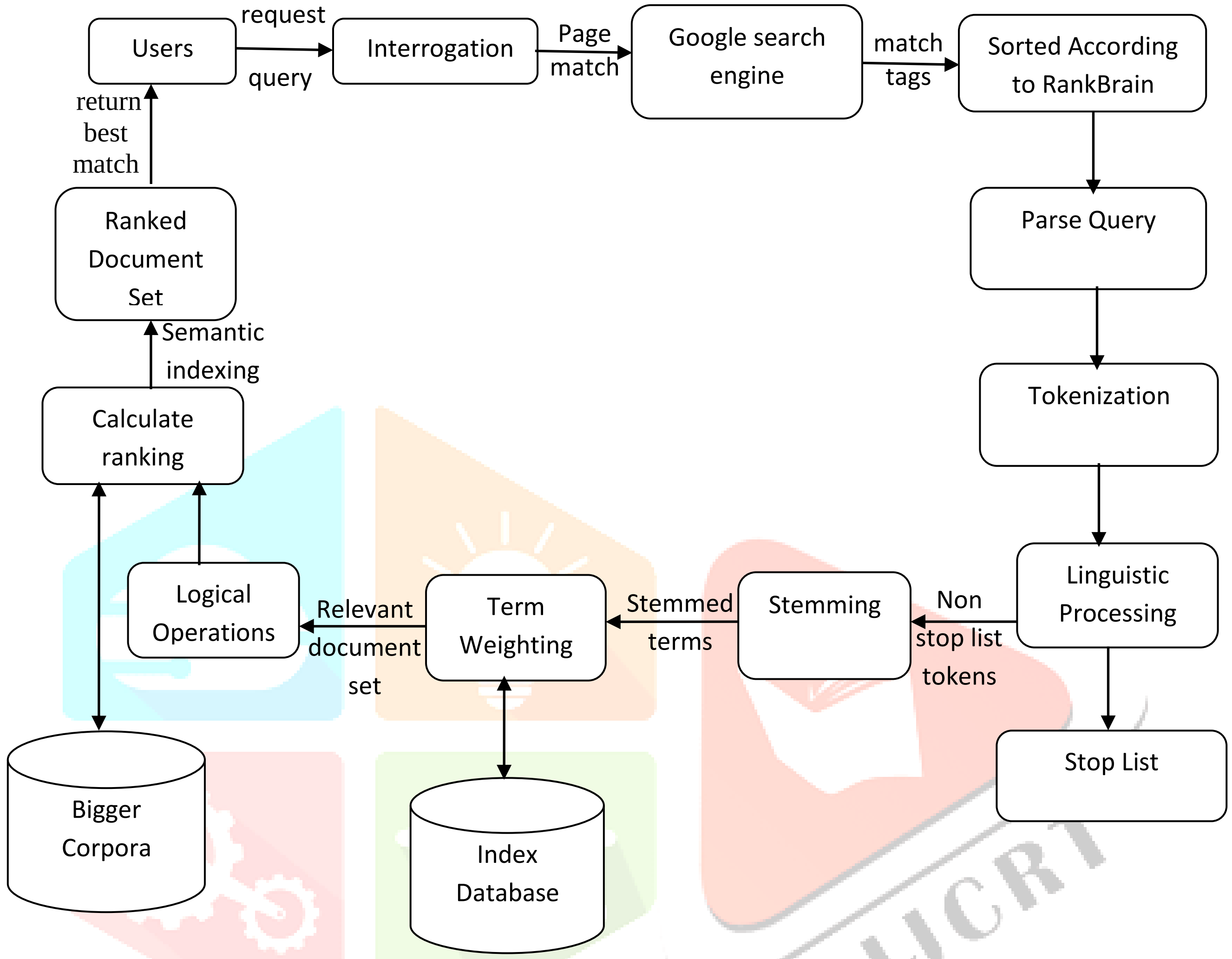

Figure1: Functional architecture of proposed framework

Explicit over existing framework

The apparent purpose of proposing framework is to provide a complete conceptual scheme for intellectual detection of the information either in form of data sets, news groups, articles etc. Such a scheme could therefore as a basis for finding similarities [10], measurement, accuracy predictions and implementation of information. The proposed framework has been designed with the perspective of an entire system, so feasible systems replace or supplement existing information processing mechanisms are required. The framework is the step-by-step procedure where user locates specific data directly or indirectly in the form of query that interrogate with search engine [15], the algorithm RankBrain arrange the page according to matching tags. Parsing of query statement is done then tokenization process is use for their unique identification. Linguistic[8] processing phase semantically[1, 2, and 18] evaluate the tokenized segment and nonstop list tokens are carry forward for stemming that refers to natural unrefined process that cut off sentence into non changing words (for example argue, argued, argues, arguing and argus the stem would be argu) otherwise pass it to stop list. After stemming, the stemmed term are measured with the collaboration of index database and generate relevant document set. Furthermore logical operation performed to literally evaluate the sentence. Using bigger corpora ranking are calculated based on their accuracy and semantically arrange in increasing order. Finally ranked document set returned to the user on the basis of best match outcome.

Proposed framework is designed in a modular fashion as illustrated in Figure1 consists of the following steps:
Step1: Inputting the service by user

Step2: Computing an Interrogation order
Step3: Making the use of Google Search Engine
Step4: Arranging the list in sorted order to screen RankBrain
Step5: Parsing the query to analyze syntactically by assigning a constituent structure
Step6: Tokenism to achieve goal
Step7: Putting a linguistic process based on prescribed procedure
Step8: Start the process of stemming
Step9: Create an Index Database
Step10: Term weighing
Step11: Perform the logical operations
Step12: Create the process of calculating RankBrain
Step13: Take the bigger Corpora
Step14: Build the calculated rank set

V. Efficiency & Applicability of the Proposed Framework

We evaluate the procedure followed by searching algorithms based on document similarity with the best matching term. We analyze various similarity searching strategies [13] e.g. content quality based, location & distance based, exact word based, rank based, semantic based etc followed by searching algorithms.

(i) One of the advantage of this approach is the relevancy to a query of the query itself in document , new methodology is able to interact with several different applications thus the model is not only a piece of design but a complex construction that can work on various data sets.
(ii)Portability between tests and graphics, one major benefits of similarity detection approach is the possibility to "do things as simple as possible". In order to maximize simplicity and this approach supports the portability among various platforms.
(iii) It is a systematic, innovative methodology, from many reasons it builds systematically that can handle complexity well. Almost covers all the factors query and retrieving direct or indirect impacts on information.
(iv) It allow for legibility. This model uses the model as the "common language" readable comprehensible to all of the experts and not just by as few detection specialist software.
We can apply to the applications where result is not based on what you see is what you get while what you see is cognitively directed to other findings. The contribution of this paper is an innovative technique for solving learning problems using similar searching [12] i.e. not yet solved by simple models.

VI. Conclusions

Thus we can conclude that our novel representation is applicable to detect similarity for document where outcome is based on semantic interpretation of complex query. The proposed framework should be able to adopt different architectures. The proposed approach uses the RankBrain Google search algorithm for searching scenario. The presented methodology does not strive for the searching of the layers of basic architectures in isolation, but to search the whole architecture. The method is based on Tokenism and linguistic process to achieve goal for the given applications. The presented approach also supports the simulation model.

References

[1] D. Hassana, M. Might, and V. Srikumar.: A Similarity-Based Machine Learning Approach for Detecting Adversarial Android Malware.: University of Utah UUCS-14-002

[2] M. Fowke, A.Hinze, R. Heese.: Text Categorization and Similarity Analysis.: Implementation and Evaluation.:Department of Computer Science The University of Waikato Private Bag 3105 Hamilton, New Zealand 2 Pingar International Ltd,2013 .

[3] H. Wael.:A Survey of Text Similarity Approaches.: GomaaComputer Science Department Modern Academy for Computer Science & Management Technology Cairo, Egypt Aly A. Fahmy Computer Science Department Faculty of Computers and Information, Cairo University Cairo, Egypt.: International Journal of Computer Applications (0975 – 8887) Volume 68– No.13, April 2013.

[4] I. FUJINO and Y. HOSHINO.:Finding Similar Tweets and Similar Users by Applying Document Similarity to Twitter Streaming Data.: (received on September 28, 2013 & accepted on January 27, 2014)

[5] S.A.Hiremath, M.S.Otari.:Plagiarism Detection-Different Methods and Their Analysis: Review .: International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163 Volume 1 Issue 7 (August 2014)

[6] Shantam , H. Kumar and Dr. RK Tiwary.:ANALYSIS OF RANKBRAIN ALGORITHM USING MACHINE LEARNING ,Department of Mechanical Engineering, RVSCET Jamshedpur, India.: International Journal of Research in Engineering, Technology and Science, Volume VII, Special Issue, Feb.2017, ISSN 2454-1915

[7] A. Chandra, M. Suaib, and Dr. R. Beg .:GOOGLE SEARCH ALGORITHM UPDATES AGAINST WEB SPAM Department of Computer Science & Engineering, Integral University, Lucknow, India. Informatics Engineering, an International Journal (IEIJ), Vol.3, No.1, March 2015

[8] S. Alzahrani, N. Salim, and A. Abraham, SMIEEE.:Understanding Plagiarism Linguistic Patterns, Textual Features and Detection Methods.: IEEE TRANSACTIONS ON SYSTE MS, MAN, AND CYBERNETICS - PART C: APPLICATIONS AND REVIEWS, VOL. xx, NO. xx, MMM 2011

[9] L. Morris, B. Myers, C. Sanders .:Implementation of Hummingbird Cipher in Python.: Department of Computer Science,Rochester Institute of Technology,Rochester, New York 14623, USA 2012/11/1

[10] F. Gao and W. Derguech .:Ubiquitous Service Capability Modeling and Similarity Based Searching.: Digital Enterprise Research Institute, National University of Ireland, Galway, Ireland firstname.lastname@deri.org, Web Information Systems Engineering – WISE 2011 and 2012 Workshops pp 173-184Lecture Notes in Computer Science book series (LNCS, volume 7652)

[11] P. P. Roy,U. Pal,J. Lladós.: Word searching in unconstrained layout using character pair
Coding.: IJDAR (2014) 17:343–358 DOI 10.1007/s10032-014-0227-6 © Springer-Verlag Berlin Heidelberg 2014

[12] P Zezula .:Similarity Searching for the Big Data Challenges and Research Objectives.: Mobile Netw Appl (2015) 20:487–496 DOI 10.1007/s11036-014-0547-2 © Springer Science+Business Media New York 2014

[13] Taher et al.:Evaluating Strategies for Similarity Search on the Web.: Stanford University Computer Science Department Stanford, Copyright is held by the author/owner(s). ªlªlª «>¬¬"«, May 7– 1,2002,Honolulu,Hawaii,USA.ACM 1-58113-449-5/02/0005

[14] C. Fries et al.: GOOGLE'S RANKBRAIN and the future of SMART SEARCH.: BLF Bigger Law Firm
Magazine 4023 Kennett Pike Suite 57516 Wilmington, DE 19807 November/December | Vol. 37 2015

[15] W. B. Croft,D. Metzler,T. Strohman.: Search Engines Information Retrieval in Practice.: book published by: Pearson Education, Inc © 2015

[16] Shai Shalev-Shwartz and Shai Ben-David .:Understanding Machine Learning:From Theory to Algorithms.: Published by Cambridge University Press ©2014 .http://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning

[17] Johanna Geiß.: Latent semantic sentence clustering for multi-document summarization.:University of Cambridge Technical Report Number 802 Computer Laboratory UCAM-CL-TR-802 ISSN 1476-2986 July 2011.

[18] E. Canhasi .: Measuring the sentence level similarity.:University of Prizren Prizren, Kosovo ISCIM , pp. 35-42 © 2013