Int. J. Nonlinear Anal. Appl.
Volume 12, Special Issue, Winter and Spring 2021, 1655-1664
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2021.5852

Optimization of web search techniques using frequency analysis

M. Aishwarya[a,*], N. Ilayaraja[a], R.M. Periakaruppan[b]

[a]Computer Applications, PSG College of Technology, Coimbatore, India
[b]Applied Mathematics and Computational Sciences, PSG College of Technology, Coimbatore, India

(Communicated by Madjid Eshaghi Gordji)

Abstract

The raw data obtained in the form of search results may be large for any particular problem, but is often a relatively small subset of the data that are relevant, and a search engine does not enable discovering the necessary subset of relevant text data in a large text collection. In this paper, a solution to a problem called conformity to truth, which studies how to find websites with the maximum amount of true facts, from a large amount of conflicting information on the user-defined topic, is proposed. Two algorithms called ParaSearch and FactFinder, which helps in identifying the best web links for searching general information and finding individual facts respectively are proposed. In ParaSearch, latent Dirichlet allocation (LDA) is used to identify the top 10 frequent terms using which we further construct a similarity matrix to identify the best web pages. In FactFinder, the usage of semantic processing is done to identify the best web pages, building upon the existing Page Rank Algorithm to further optimize the search results. The results prove that ParaSearch can identify web pages with the maximum number of facts conforming to the truth much better than popular search engines. The ambiguity of the individual facts is decreased to a great extent by using the FactFinder algorithm. Thus these algorithms will increase the accuracy of identifying possible web links for a given search word much better than most of the popular search engines.

Keywords: frequency analysis, latent Dirichlet allocation, text mining.

*Corresponding author
Email addresses: aishwarya.prabhamurali@gmail.com (M. Aishwarya), nir.mca@psgtech.ac.in (N. Ilayaraja), rmp.amcs@psgtech.ac.in (R.M. Periakaruppan)

Received: August 2021 *Accepted:* September 2021

Table 1: Conflicting data about height of Mount Everest

Web Site	Height
Wikipedia.org	29,029 ft
History.com	29,002 ft
Britannica.com	29,035 ft
Thedailybeast.com	26,000 ft
Independent.co.uk	29,029 ft

1. Introduction

Text data are unique in that they are usually generated directly by humans rather than a computer system or sensors, and are thus especially valuable for discovering knowledge. The World Wide Web has a plethora of information both relevant and irrelevant to the subject being searched. For example, while searching for the word "raagas", people find results from a variety of pages ranging from the music site "raaga.com" to social networking profiles of people with last names matching the keyword, to some sites pertaining to the meaning of raaga.

Another case of information mismatch occurs while trying to find particular facts pertaining to a subject.

Example 1.1. ***Height of Mount Everest.*** *While trying to ascertain the height of Mount Everest, we gathered varying values for the height. The conflicting information is projected below in table 1.*

Referring a web page is done to obtain information about a certain subject. But when the web pages provide multiple pieces of information, how measure the correctness of it? Hence a solution for conformity to truth is presented, wherein the identification of the data which has the maximum probability of being the correct information is done.

Many algorithms like page rank algorithm [12], Authority hub analysis [10] or other general link based algorithm [4] associate popular web pages pertaining to the subject. But the responsibility of discerning true facts from false facts is left undone.

This paper describes the conformity to truth problem and two algorithms FactFinder and ParaSearch are proposed along with its underlying framework, to eliminate the problem described to the maximum extent possible. These algorithms are more efficient in retrieving accurate information compared to the existing Page Rank.

The rest of the paper is organized as follows. Section 2 briefs about the related works in this field. Section 3 gives a brief overview of conformity to truth problem; Section 4 presents ParaSearch and FactFinder, the proposed algorithms to increase the probability of identifying true and relevant information. Section 5 presents the analysis of the algorithms. Finally Section 6 concludes the paper with future work and additional comments.

2. Related Works

Google's Page Rank [12] as well as the authority hub analysis [10] algorithms utilizes the hyperlinks to find the pages with high authorities. In page rank, the popularity of a page or the weight of a page is dependent on the number of its outbound links. A page having a large number of outbound links may not necessarily contain accurate information. In Authority hub analysis, the scope of the algorithm is limited only to broad topic queries like discerning some general information about a topic. Specific topic queries that detect some particular facts about a topic cannot be handled by the algorithm. The proposed system overcomes the disadvantages of both these systems.

3. Conformity to Truth

Inconsistent information among different websites indicates potential data quality problems such as accuracy, completeness, timeliness, etc. This can mislead the user about the credibility of the information and will prevent the effective use of information [1]. To solve the problem of conformity to truth, it is assumed that the same information that is presented by multiple web sites has a much higher probability of being reliable. On this assumption FactFinder and ParaSearch algorithms try and identify websites whose information concurs with each other. The relevant original text data are examined to identify the relevant subset of data. To this subset both semantic and non-semantic methods are applied to mine data.

4. Proposed Algorithm

In order to improve the search results for both semantic and non-semantic query search in web sites Factfinder and Parasearch Algorithms are proposed and experiments have been conducted to show our algorithm out performs while returning search results.

4.1. Factfinder and Parasearch Algorithms

Both ParaSearch and FactFinder algorithms utilize the search results provided by Google (through its API). Based on whether the user wants to search for a fact or some generic information about the subject, the results obtained from Google are passed to FactFinder and ParaSearch algorithms respectively. These algorithms further refine the search so as to increase the probability of finding true data pertaining to the subject.

4.2. ParaSearch Algorithm

The main utilization of ParaSearch algorithm is to retrieve links with maximum information pertaining to the subject. This algorithm also aims to maximize the veracity of the individual information in each link. We try to identify the web pages, which has a large number of facts or sentences pertaining to the topic to be trustworthy and each sentence is claimed to be trustworthy if it is present in multiple web pages. Xiaoxin yin et al [6] showed that the truth finder assumption gives better results.

The methodology used in the algorithm consists of first obtaining the search results from Google (through its API) and using the links obtained, the corresponding text files are retrieved. To the text files obtained, we perform data cleaning in the form of stop words removal to prevent any stop words showing up as results in the topic wise key words obtained on applying LDA [3]. As discussed by Eduard et al [5], stop words are words which do not convey any significant semantics to the texts or phrases they appear in. Latent Dirichlet allocation is a generative probabilistic model for document modeling and text classification. In the context of text modeling, the topic probabilities provide an explicit representation of a document.

Latent Dirichlet allocation is applied to the text documents. For a given documents, LDA can identify θ number of topics. For each of the topics identified, it can identify Φ number of words which indicate the topic. Hence the θ value is set, which is indicative of the number of topics to be identified from each document as 1. The Φ value which indicates the number of words per topic is set as 10. By applying LDA, the central topic of discussion of each web page is identified and for each of these topics, the top 10 keywords which act as indicators of the topic are also identified.

---

www.google.com; gson 2.2.1

Table 2: Sample term comparison matrix for ParaSearch algorithm for the term Himalayas

	Himalayas	mount	...	range
Document1 https://en.wikipedia.org/wiki/Himalayas	1	1		1

A term frequency matrix or a tf [2] is used as a means to measure information retrieval. A term frequency matrix or document term matrix is a matrix consisting of terms as column headings and documents as the row headings. Here the matrix is formed with the collaboration of all the key words obtained in the procedure as the column headings and in the text documents of the search results as the row headings.

Here the matrix values are considered as being binary valued, i.e. 1 if the term is present in the document, else 0. After the construction of the matrix, the row wise counts are calculated to check which document has the maximum number of terms present. Such documents have a high probability of having facts related to the search term. The equation for the count is given by (4.1).

$$Count_1 = \sum_{i=1}^{k} topic\ index[i] \tag{4.1}$$

where k is the total number of terms identified from all the documents combined. Here we consider the matrix column to start from the indexing of 1, hence we start with $i = 1$.

By using topic word distribution of each link, an attempt is made to maintain the veracity of the facts, provided by the link. The links with maximum count value indicates that it has matched the maximum number of attributes (key words). This in turn is an indication of the fact that the majority of the information is present in the link. The output will be the top links with the highest count values in descending order.

Here in table 2, the topic indicator words of the searched word are shown. The stepwise algorithm for ParaSearch is as follows,

Step 1 Read the input.

Step 2 Based on the searched words (inputs) search for the related web pages.

Step 3 Based on the Page Rank algorithm retrieve top ten web page links.

Step 4 in the extracted sentences of each link,

1. Remove the whitespaces
2. Remove the tab spaces
3. Remove the punctuation marks
4. Remove the stop words(pronouns, conjunctions, etc.)
5. Transform the characters to lower case

Step 5 We will obtain only the keywords in each link. Apply Latent Dirichlet allocation to obtain the top 10 keywords.

Step 6 Using the keywords, form a dissimilarity matrix for all the links.

Step 7 The dissimilarity matrix is formed in such a way that links are arranged row wise and the key words are arranged as attributes.

Step 8 The dissimilarity matrix is obtained by comparing each word in a link with other words in all other links.

Step 9 Maintain a count in each row.

Step 10 Select five links which have highest count.

The number of pages considered from Google is the first ten results and the subsequent steps of the ParaSearch algorithm are performed on these results. Finally the top five results in the decreasing order of their count values obtained using the term comparison matrix are given as results.

4.3. FactFinder Algorithm

The main utility of the fact finder also lies in identifying a fact defined by the subject. Natural language processing is used to obtain the required results. Fact finder algorithm is used if the user wants to identify a particular fact about some topic. For e.g.: What is the height of Mount Everest? Here the user wants to ascertain only the exact height value of the mountain. In such cases, this algorithm is used.

As in the previous section, the links are retrieved through Google (through its API), and subsequently the text contents are also retrieved. Here the assumption that a sentence having the maximum number of queried key words as a high probability is used. So to identify the maximum of key words, a simple function x/2, where x denotes the number of queried words is used. If a sentence has a match, it is selected, else rejected.

In some cases, owing to the presence of the personal pronouns, the need to make allowances in the searching arises. So the next sentence to the one previously retrieved is also considered. For example, Mount Everest also known as Sagarmatha and in Tibet as Chomolungma, is Earth's highest mountain. Its peak is 8,848 meters above sea level.

In the above example, it can be noted that the actual value of the height is not given in the first statement but in the second sentence, where the words Mount Everest is not present, but indicated by the presence of a pronoun "it". Due to this line of reasoning, both sentences containing the key words and the next sentence are retrieved for the analysis. As with the previous algorithm, the stop words from all the sentences so far retrieved are removed. As the stop words are basically conjunctions, pronouns or prepositions, the quality of the result is not affected [5].

The user searched facts will be a subject or an object and that in turn will either be a noun or an adjective. As defined by Jurafsky, "noun includes the words for most people, places or things". Adjectives are "a class that includes many terms for properties or qualities". Based on these, nouns and adjectives can be extracted from the sentences so far retrieved using natural language processing. The term comparison matrix is formed using these nouns and adjectives, obtained from all the documents as the column headings and documents as the row headings. Same as in the previous algorithm the matrix values are binary with 1 if this term is present in the document and 0 if it is absent. The final count of each row is obtained as per equation (4.2) given as follows:

$$Count_2 = \sum_{i=1}^{k} matrix[i] \tag{4.2}$$

Where k denotes the total number of nouns and adjectives found from all the documents. Here we consider the matrix column to start from the indexing of 1, hence we start with $i = 1$.

Duplicate values are deleted from the column heading. The sample matrix format obtained with just the first page result while searching for "height of Mount Everest" follows fig 2.

Table 3: Format of term comparison matrix for FactFinder algorithm

	Mount Everest	Sagarmatha	Chomolungma	8,848
Document1 https://en.wikipedia.org/wiki/Mount_Everest				

The number of nouns and adjectives may be different in different documents. The documents having the highest count value in descending order is identified and the links are displayed to the user. Here, by comparing with different websites, if a fact is present in multiple websites, it is considered as being true. Hence those web sites are selected as being trust worthy with regard to the facts being searched. The algorithm for FactFinder is given in a step wise manner,

Step 1 Read the input.

Step 2 Based on the searched words (input) search for the related web pages.

Step 3 Based on the Page Rank algorithm retrieve top ten web page links.

Step 4 Retrieve the sentence containing the words searched from each link.

Step 5 In the extracted sentences of each link,

1. Remove the whitespaces
2. Remove the tab spaces
3. Remove the punctuation marks
4. Remove the stop words(pronouns, conjunctions, etc.)
5. Transform the characters to lower case

Step 6 We will obtain only the nouns and adjectives in each link using NLP. Using that, form dissimilarity matrix.

Step 7 The dissimilarity matrix is formed in such a way that links are arranged row wise and the key words are arranged as attributes.

Step 8 The dissimilarity matrix is obtained by comparing each word in a link with other words in all other links.

Step 9 Maintain a count in each row.

Step 10 Select five links which have highest count.

The number of pages considered from Google is the first ten results and the subsequent steps of the FactFinder algorithm are performed on these results. Finally the top five results in the decreasing order of their count values obtained using the term comparison matrix are given as results.

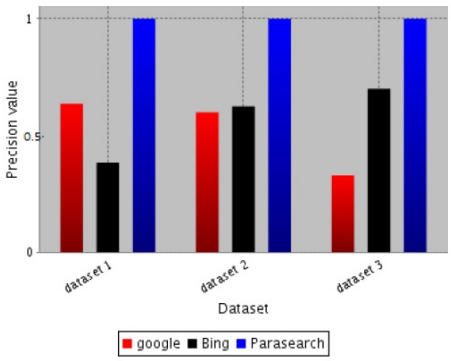

Figure 1: Precision measure of Bing vs Google vs ParaSearch

5. Experimental Results and Analysis

In this section, experiments which show the effectiveness of the algorithms are presented. The measures of precision and recall are used to measure the effectiveness of ParaSearch and FactFinder. The searches are compared against the existing Google search and Bing search.
Precision and recall are the basic measures used in evaluating search strategies. Precision is the ratio [7] of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. Recall is the ratio of relevant records retrieved to the total number of relevant records in the data base. There exists an inverse relation between precision and recall.
A large number of terms using ParaSearch [8], Google [9, 10] and Bing [11] are searched and the results are obtained. When the graph was plotted for both precision and recall, a distinct increase in the performance of the proposed ParaSearch algorithm was found.
When general information about Himalayas was searched, the Google search results in its first page had 8 results, out of which only 5 were pertaining to the Himalayas. One was about Himalaya's herbals, the other one was about trekking packages, and the last about Himalayan cosmetics. When, the same keyword was searched in Bing, 8 results were found, with only 4 results giving out actual facts about the mountain, while 1 was a dictionary definition of the word, 3 of them were about trekking and travel party. This same keyword when searched using ParaSearch, 4 results were found and all 5 were pertaining to the information about the mountain ranges.
This method was repeated in random using Google, Bing and ParaSearch, a distinctive improvement was found in the performance of the proposed FactFinder algorithm as against Google and Bing. When comparing ParaSearch, Bing and Google, ParaSearch has an increase of 76% precision over Google search engine and 36% over Bing search engine and a decrease of 24% recall over Google and 32% decrease over Bing as indicated in figure 1. Precision and recall have an inverse relationship, so the increase in precision may have prompted the decrease in recall values shown in figure 2.

While searching for the height of the Mount Everest using Google, Bing and FactFinder, the height values obtained are tabulated below 4.

In some searches the first page had only less number of results [12, 13], hence $*$ is used to fill the gaps in table 4.From the above tabulated values, it can be noted that the fact finder algorithm performed better than the Google's page rank or Bing search. On searching using the three algorithms with different queries, the results were tabulated, and the graphs subsequently obtained for precision or recall values show a substantial increase in the performance of the FactFinder. When comparing FactFinder, Bing and Google, FactFinder has an increase of 31% precision over Google search engine and 10% over Bing search engine and a decrease of 42% recall over Google and 43% decrease over Bing as indicated in figure 3.

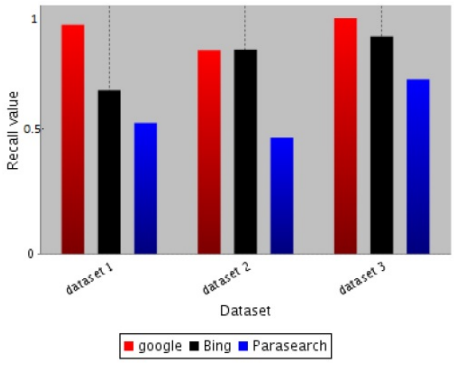

Figure 2: Recall measure of Bing vs Google vs ParaSearch

Table 4: Format of term comparison matrix for FactFinder algorithm

Google	Bing	FactFinder
29029	29000	29029
29002	29029	29029
29029	29035	29029
26000	27000	29029
29029	8850	29029
29029	29029	*
*	28000	*

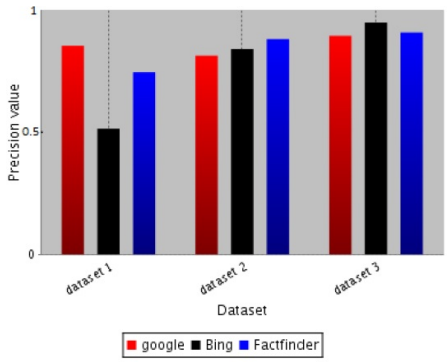

Figure 3: Precision measure of Bing vs Google vs FactFinder

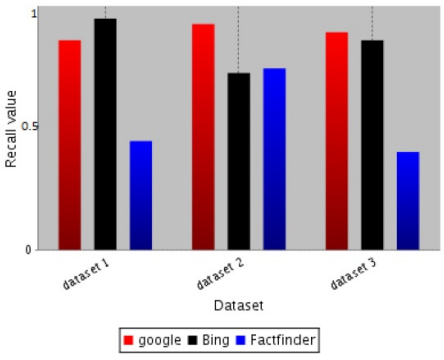

Figure 4: Recall measure of Bing vs Google vs FactFinder

6. Conclusion

In this paper, the conformity to truth problem, which aims at resolving conflicting facts from multiple websites and finding the true facts among them is defined. FactFinder algorithm is proposed which utilizes the natural dependencies of the language to identify the facts. ParaSearch is an approach that utilizes Latent Dirichlet allocation to find the topics of the related pages. Experiments show that both FactFinder and ParaSearch achieve higher precision as compared to leading search engines.

References

[1] B. Amento, L Terveen and W. Hill, *Does Authority mean quality? predicting expert quality ratings of web documents*, Proc. ACM SIGIR '00, Assoc. Comput. Machin. (2000) 296–303.
[2] I. Antonellis and E. Gallopoulos, *Exploring term-document matrices from matrix models in text mining*, Proc. SIAM Text Mining Workshop, 6th SIAM SDM Conference, Maryland, 2006.
[3] D.M. Blei, A.Y. Ng and M.I. Jordan, *Latent dirichlet allocation*, J. Mach. Learn. Res. 3 (2003) 993–1022.
[4] A. Borodin, G.O. Roberts, J.S. Rosenthal and P. Tsaparas, *Link analysis ranking: Algorithms, theory, and experiments*, ACM Trans. Internet Technol. 5(1) (2005) 231–297.
[5] E. Dragut, F. Fang, P. Sistla, C. Yu and W. Meng, *Stop word and related problems in web interface integration*, Proc. VLDB Endow. (2009) 349–360.
[6] J. Han, X. Yin and P.S. Yu, *Truth discovery with multiple conflicting information providers on the web*, IEEE Trans. Knowledge Data Engin.20(6) (2008) 796–808.
[7] R. Jizba, *Measuring Search Effectiveness*, *https* : *//www.creighton.edu/fileadmin/user/HSL/docs/ref/Searching_−_Recall_Precision.pdf*, (2007).
[8] D. Jurafsky and J.H. Martin, *Speech and language processing: An introduction to natural language processing*, Computational Linguistics, and Speech Recognition, Prentice Hall PTR, Upper Saddle River, NJ, 2000.
[9] Y.A. Kim, M.T. Le, H.W. Lauw, E.P. Lim, H. Liu and J. Srivastava, *Building a web of trust without explicit trust ratings*, 2008 IEEE 24th Int. Conf. Data Engin. Workshop, (2008) 531–536.
[10] J. M. Kleinberg, *Authoritative sources in a hyperlinked environment*, J. ACM 46(5) (1999) 604–632.
[11] Y. Matsuo and M. Ishizuka, *Keyword extraction from a single document using word co-occurrence statistical information*, Int. J. Artificial Intell. Tools 13(01) (2004) 157–169.
[12] L. Page, S. Brin, R. Motwani and T. Winograd, *The Pagerank Citation Ranking: Bringing Order to the Web*, Technical Report, Stanford Digital Library Technologies Project, (1999).
[13] S. Ramachandran, S. Paulraj, S. Joseph and V. Ramaraj, *Enhanced trustworthy and high-quality information retrieval system for web search engines*, Int. J. Comput. Sci. 5 (2009).