Optimizing Performance of User Web Browsing Search

Sunita[1](✉) and Vijay Rana[2](✉)

[1] Arni University, Kangra, India
sunitamahajan2603@gmail.com
[2] SBBS University, Khiala, India
vijayrana93@gmail.com

Abstract. Web crawling and word sensing are critical nowadays. In case of web browsing, searching consume time in case proposer requirements from user is not extracted. In earlier work on web browsers word correction was missing which is a main inclusion in the proposed work. The problem with existing literature is time complexity in fetching the correct keyword from user query string. We propose character shuffle pre-processing searching mechanism. Using the proposed method, time complexity is reduced since clustering is used for searching the keywords. The searching don't required entire database to be searched over rather only particular cluster is searched. To fetch meaningful keywords database is maintained. The keywords within the database increases as more and more user interact with this search engine. The worth of this study is proved using parameters execution time and number of meaningful keywords.

Keywords: Preprocessing · Clustering · Time complexity · Word sensing

1 Introduction

The detecting word sensing is critical and complicated task. The techniques following as under category is known as Word Sense Disambiguation (WSD). Is follow with NLP with reduced the energy consumption. The phases with natural language processing (NLP) is include with preprocessing, feature extraction, segmentation and classification. Preprocessing indicate removing any abnormal present in the data. The feature ext is next phase, which is used in order to fetch the critical necessary feature out of the available information. The segmentation is a phase which is used to divide the information into critical and non-critical words. Classification is the last phase which is used to divide the information into correct phase all of these phases area critical in general data mining. NLP (Natural language processing) uses this specific field to extract the meaningful information out of the user query.

The problem associated with word sensing for all under NP hard, NP hard problem is complex problem since a same word has contain different meaning associated from user query. They consider the two sentences underneath

A. K. Luhach et al. (Eds.): ICAICR 2018, CCIS 956, pp. 230–239, 2019.
https://doi.org/10.1007/978-981-13-3143-5_20

E.g.

(1) “I am sit near the bank”.
(2) “What is interest of the SBI bank”.

The word bank obviously has different meanings in the two contexts above [13]. In the primary first context it implies the bank of the river and in the second it implies the money of bank. The machine can’t to find the actual meaning of the words. In this case to be need a trained the system to extract the sense of the words. There are four regular ways to deal with Word Sense Disambiguation

- **Unsupervised methods:** Unsupervised [1] models concentrate on taking in an example in the information with no outside input. Clustering is an exemplary case of unsupervised learning model.
- **Semi-supervised methods:** Semi-Supervised learning [2] uses a set of curated, labelled data and tries to infer new labels/attributes on new data sets. Semi-Supervised learning models are a solid middle ground between supervised and unsupervised models.
- **Supervised methods:** Supervised learning [3] models use external feedback to learning functions that map inputs to output observations. In those models the external environment acts as a “teacher” of the AI algorithms. These use word sensing methods to learn from labelled preparation sets. A number of the general techniques used are “decision-lists”, “decision trees”, “naïve-Bayes”, “neural-networks”, “support vector machines” (SVM).
- **Knowledge-based methods:** Reinforcement learning models use opposite dynamics such as rewards and punishment to “reinforce” different types of knowledge. This type of learning technique is becoming really popular in modern AI solutions. Knowledge-based methods rely on “dictionaries-based”, “thesauri” and “lexical-resources” for knowledge bases.

In unsupervised learning dictionary maintains is not possible since customization is not possible. Semi-supervised this mechanism could be time consuming is expensive and it could be parsley customized. This is mechanism historical data plays a part the proposed work uses present and future work associated searching. Hence it can’t be used along with proposed system. Supervised learning it is customizable and can be used along with proposed.

The learning mechanisms greatly influence the pattern by which discovery of normal and abnormal phrase is made. For this purpose, supervised learning mechanism is proposed in this research. The word correction and searching take into consideration application program interface (API) from online source JOC. Survey of the article is organised as below: part 2 gives the literature review, part 3 illustrated the gaps, part 4 recently the proposed system, part 5 gives the results and last details are gives in the conclusion and future scope.

2 Literature Survey

The literature is conducted to look for optimal technique used for browsing websites with minimum amount of time consumed.

[4] proposed model suggested a social content unfolding along the line of semantics and time. Clustering [5] mechanism is imposed reducing the overall search time required. [6, 7] proposed a challenging task of surveying through the mechanism used for sentiment analysis. Sentiment analysis techniques suggested in the literature used to accurately predict the desire of the user by looking at the search query. [8] In this article highlights many searching techniques with various searching algorithms like fast string search algorithm with vector approach & bi linear search. [9] The proposed method beats different baselines and before proposed web-construct semantic closeness measures in light of three benchmark datasets demonstrating a high relationship with human appraisals. Proposed strategy fundamentally enhances the precision in a network mining assignment. [10, 11] Proposed framework a lexical example extraction algorithm to extricate various semantic relations that exist between two words. Work is led on datasets of vague questions, demonstrate that our approach enhances query output clustering as far as both clustering quality and level of expansion. [12] Proposed a portion based KNN clustering algorithm which enhanced exactness of KNN clustering algorithm. The proposed algorithm KKNNC utilizing the six UCI data sets, and contrasted it and KNNC algorithm in the tests. The exploratory outcomes demonstrate that KKNNC algorithm outflank KNNC algorithm in precision fundamentally. [13] Proposed model to distinguish some ease of use related issues in Semantic Web. Ease of use of some catchphrase and shape based instruments and their restrictions are being talked about. Results and discoveries of an ease of use study of the device are exhibited. [3] Proposed framework examines different systems for client driven relationship of inquiry and thinking. Human critical thinking in psychological science, a client inquiry in view of connected client interests. The multi-level technique from the human critical thinking to vast scale look was proposed through [3]. [14] proposed exponential law-based intrigue's safeguarding demonstrating, arrange statistics– based data gathering, and philosophy managed various leveled thinking were created to execute client question parsing and seeking criteria. Talked about methods utilized for question expectation recognition, exploiting client conduct to comprehend their interests and inclinations on web based business. Technique was intended for utilizing the substance of internet searcher result pages (SERPs), alongside the data acquired from question strings, to examine qualities of inquiry aim, with a specific spotlight on supported pursuit.

In the studied literature, execution time is sufficiently high due to lack of clustering and redundant information search and retrieval. The proposed system utilizes the tokenization and keyword searching mechanism for effectively finding the resources required for user query.

3 Research Gap

The existing literature provides content based searching however does not eliminate redundant keywords. Also dissimilar keyword searching and elimination is missing causing higher execution time and least efficient URL retrieval. This system will require large amount of information in order to make correct decision. The information which is provided to the recommender system must be consistent in nature. For the information some sort of information system is required. The recommender system will take the information and formulate the decision in one of the following two ways-either by the use of collaborative filtering or by the use of content filtering. The collaborative filtering is the mechanism of filtering for information among the multi agents, viewpoints, data sources etc. The content filtering on the other hand is the mechanism of using the program in order to filter the information which is going to be used within the system. People now days are more and more concerned with the environment. For this purpose concise information retrieval system is required.

For this purpose, efficient parsing and correction system along with clustering for reducing execution time is designed. The proposed model is described in next section as.

4 Proposed Model

Proposed model combination of multiple phases and Parsing is one of the critical phases.

- **Parsing**

Extracting the meaningful information out of the particular string is the main objective of the parsing. In order to do this, space is act as the separator. Example "My name is Sunita Mahajan". Suppose we have a dict.mdb database.

Since the specified words matched with dictionary hence successful tokenization & as well as parsing is done. "my", "name", "is", "Sunita", "Mahajan". After performing parsing successfully extract the meaningful keywords from the given string.

Table 1. Showing the dictionary containing words along with meaning

Words	Meanings
My	Personal
Name	Title, label
Is	Am, are
Sunita	Daughter of Dharma, good behaviour
Mahajan	Castes, communities

- **Finding meaningful keywords**

Another dictionary with the co-related words is maintained order to determined meaning of the sentence. The match is counted as hit and no match is indicated with missed. The main task of our approach is to increase hit and missed occur words replaced with corrected words. The equation is used to calculate hit to miss ratio.

$$TS\_hit\,ratio = \frac{Hits_i}{TS_i} \quad (1)$$

Equation 1: Total hit ratio

This equation indicates the total number of keywords fetched by the proposed system to the total keywords present within the dictionary. The result is presented in the form of percentages.

In the word sensing model which is proposed the hit ratio is given by considering the total words count of 100 in a dictionary.

Table 2. Comparison in terms of ratio

String searched	Hit ratio with existing system (ontology based model)	Hit ratio with proposed system (user perception based)
Live cricket score today	0.03	0.04
I am sit near the bank of the river	0.08	0.09
What is the interest of the current in SBI bank	0.08	0.10
SBI saving account interest rate per month	0.06	0.07
Fashion is a popular style, especially in clothing, footwear on Amazon	0.10	0.11

The hit ration ex and pro indicated that the result of pro model is better since the words which does not exists in the dictionary are added to the dictionary with user permission. This procedure of higher hit ratio as compared to existing model (Fig. 1).

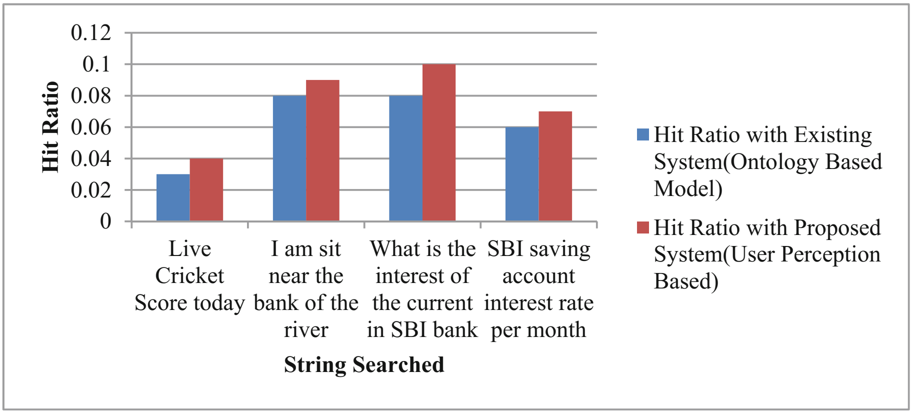

Fig. 1. Comparison of hit ration of existing system & proposed system

4.1 Proposed Algorithm

The algorithm which will describes the creation of Recommender system for the promotion of Selected Websites is describes through the following steps.

```
Algorithm-Auto Query Resolving system
a. Receive the parameters of the user query to be tested(Pi)
b. Divide the string into meaningful tokens (t) is also known as parsing.
•  In the parsing space is act as the separator.
•  The (t) are matched within dict.mdb (db) database.
•  Extract the meaningful keywords (k).
c. Find the sense(s) of the k.
•  The meaning of the words are found using hit ratio
•  hit ratio = $\frac{\mathrm{Hits_i}}{\mathrm{TS_i}}$
d. check for availability
•  if(t==db) is rejected
•  then
•  Otherwise it will be added into (db).
•  Results as present into percentage
e. Stop
```

In the proposed algorithm first step receive the parameters of the user query to be tested (Pi). In the second step preprocessing, divide the string into tokens, this process is also known as parsing. Parsing act as a separator after the parsing extract the keywords are matched within a dict.mdb database. Find the meaningful keywords. The actual keywords the fetched from the user query. In the next step find the ambiguous words, and find the actual sense the keywords.

The success of the system will be determined using Hit Ratio.

$$\boldsymbol{hit\,ratio} = \frac{\boldsymbol{Hits_i}}{\boldsymbol{TS_i}} \tag{2}$$

Higher the hit ratio more successful will be the given system.

The existing approach doesn’t consider the identification of similar words & also the token matching process is slow in the proposed literature the explanations variation of keyword fetch and matching consider. The complexity of the search is reducing greatly by the use of proposed system.

5 Performance and Result Evaluation

In this section performance can be evaluate on number of websites compare along with execution time. It is total time consumed to relevant websites.

- **Recall**

$$\boldsymbol{Recall} = \frac{\boldsymbol{Number\,of\,RW}}{\boldsymbol{RW + NRW}} \times \mathbf{100} \tag{3}$$

The overall performance is described in Tables 1 and 2 highlights list of four keywords with their recall measurer that describes the relevant and non-relevant result.

- **Single-keyword based query**

(See Figs. 2, 3 and Table 3).

Table 3. Study of quantitative analysis

Keywords	Total no. of websites retrieved	Time consumed (milli seconds)
University	9	3
Online	11	4
Computer	10	3
School	7	4

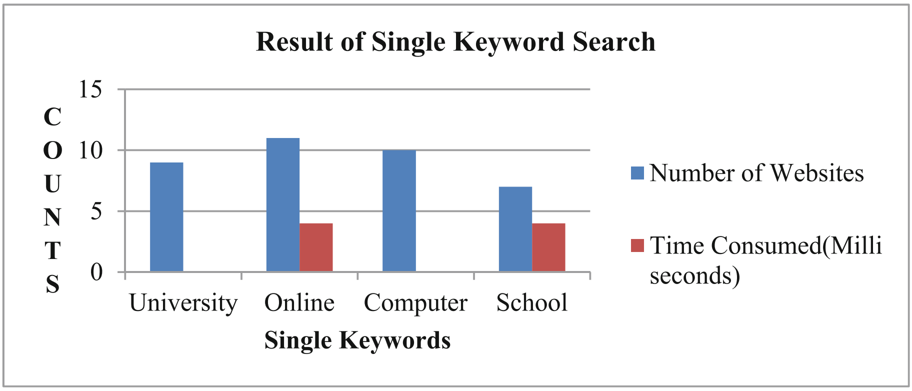

Fig. 2. Plot of frequency vs keywords

Table 4. Confusion matrix for single-keyword

Search query	No. of sites retrieved	RW (related websites)	LRW (less related websites)	NRW (non-related websites)
University	9	7	1	1
Online	11	8	3	0
Computer	10	7	2	1
School	7	5	1	1

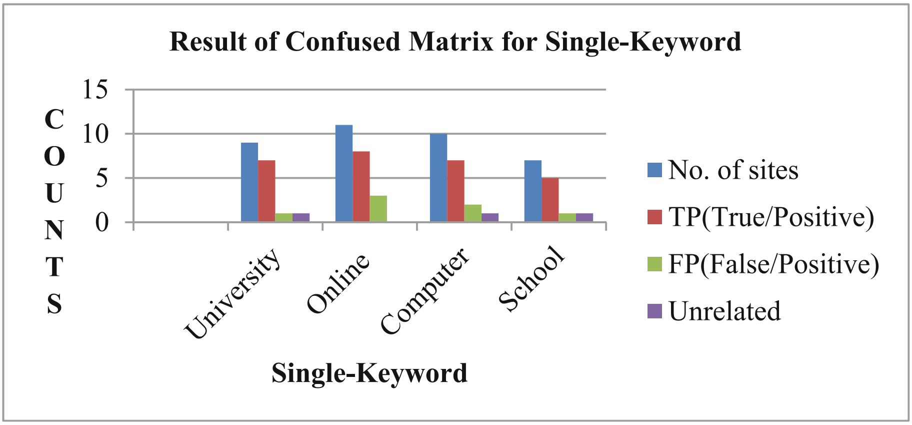

Fig. 3. Confused matrix for single keyword

- **Multi-keyword based query**

(See Figs. 4, 5, Tables 4, 5 and 6).

Table 5. Study of quantitative analysis

Multi-keywords	Total of no. of websites retrieved	Time consumed (milli seconds)
Arni University	12	5
Online shopping	11	4
Computer education	12	3
School library	11	4

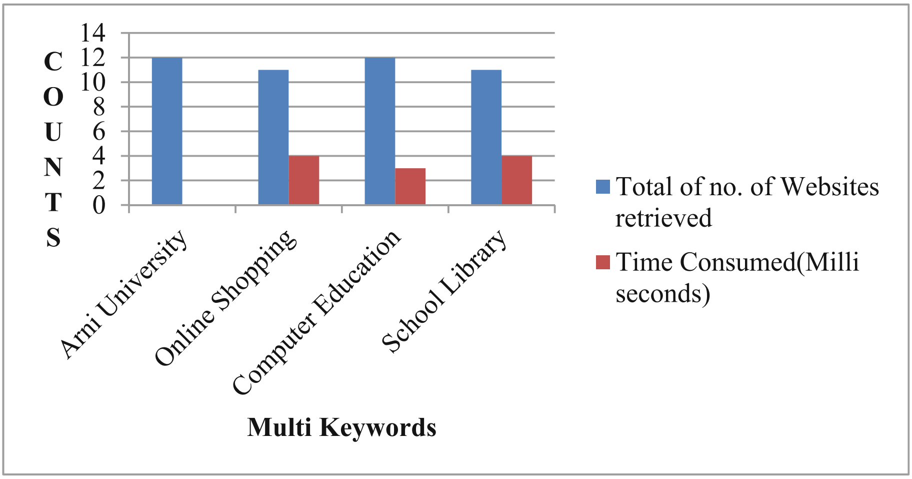

Fig. 4. Plot of frequency vs keywords

Table 6. Confusion matrix for multi-keywords

Search query	No. of sites retrieved	RW (related websites)	LRW (less related websites)	NRW (non-related websites)
Arni University	9	5	3	1
Online shopping	11	8	2	1
Computer education	12	11	1	0
School library	11	11	0	0

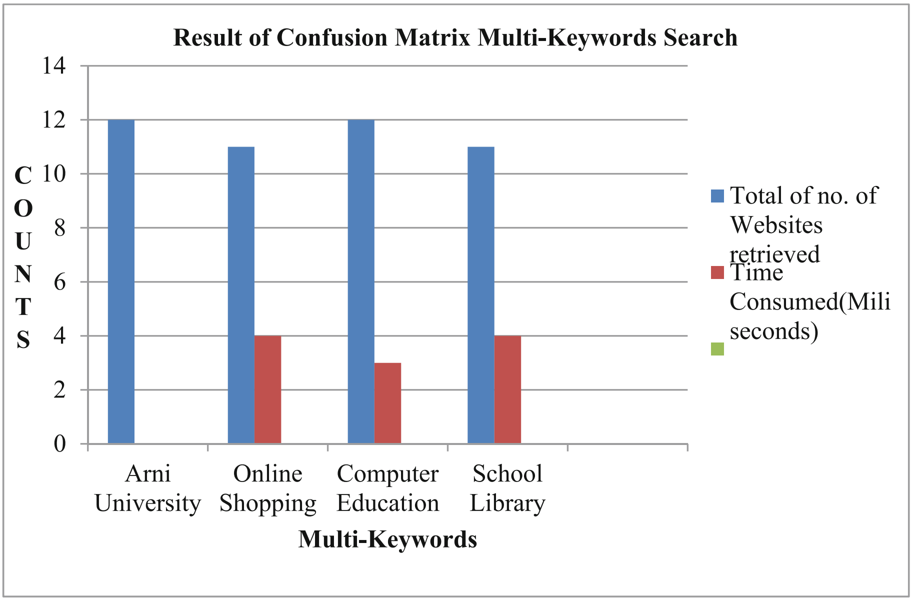

Fig. 5. Confusion matrix for multi-keywords

6 Conclusion

The result from the proposed system indicates betterment in terms of confusion matrix. The keywords matching and parsing process gives unique labels along with high precision. The keyword matching frequency yield the order at which the obtained website is going to be displayed at the browser. The pre-processing phase also filters the information to be displayed to the user keeping in mind the user interest. The time consumption in fetching the website greatly depends upon the server caches and processor speed. The proposed system is tested on single CPU but may perform better on GPU. In the future work, clustering along with sense annotation and location sensitive along with proposed system.

References

1. Che, W., Liu, T.: Using word sense disambiguation for semantic role labeling. In: 2010 4th International Universal Communication Symposium, pp. 167–174 (2010)
2. Features, S.: A method for word sense disambiguation combining context semantic features, pp. 283–287 (2016)
3. Zeng, Y., et al.: User-centric query refinement and processing using granularity-based strategies. Knowl. Inf. Syst. **27**(3), 419–450 (2011)
4. De Maio, C., Fenza, G., Loia, V., Orciuoli, F.: Unfolding social content evolution along time and semantics. Future Gener. Comput. Syst. **66**, 146–159 (2017)
5. Shekarpour, S., Marx, E.: RQUERY : rewriting natural language queries on knowledge graphs to alleviate the vocabulary mismatch problem (2017)
6. Mohey, D., Hussein, E.M.: A survey on sentiment analysis challenges. J. King Saud Univ. - Eng. Sci. **30**, 330–338 (2016)
7. Mahajan, S., Sharma, S., Rana, V.: Design a perception based semantics model for knowledge extraction. Int. J. Comput. Intell. Res. **13**(6), 1547–1556 (2017)
8. Chandra, S.: A brief study and analysis of different searching algorithms (2017)
9. Bollegala, D., Matsuo, Y., Ishizuka, M.: A web search engine-based approach to measure semantic similarity between words. IEEE Trans. Knowl. Data Eng. **23**(7), 977–990 (2011)
10. Navigli, R., Crisafulli, G.: Inducing word senses to improve web search result clustering. Computational Linguistics, pp. 116–126 (2010)
11. Rana, V.: An approaches & comprehensive survey for measuring semantic relatedness with knowledge resources, vol. 6, no. 1, pp. 398–403 (2018)
12. Wang, Y.: K-nearest neighbor clustering algorithm based on kernel methods, pp. 0–3 (2010)
13. Haider, A., Raza, A.: Keyword and form based semantic search tools and their usability. In: 8th International Conference on Digital Information Management ICDIM 2013, pp. 85–89 (2013)
14. Ashkan, A., Clarke, C.L.A.: Impact of query intent and search context on clickthrough behavior in sponsored search. Knowl. Inf. Syst. **34**(2), 425–452 (2013)