IDENTIFYING AMBIGUOUS QUERIES METHOD BASED ON SPATIAL INFORMATION FOR IMPROVING WEB INFORMATION RETRIEVAL

Shahid Kamal[1],*, Adnan Ahmed[2], Sohail Khokhar[3]
[1]Institute of Computing & Information Technology, Gomal University, D.I.Khan, KPK, Pakistan
[2]Department of Computer Systems, Quaid-e-Awam University of Engineering, Science and Technology, Nawabshah, 67450, Pakistan
[3]Department of Electrical Engineering, Quaid-e-Awam University of Engineering, Science and Technology, Nawabshah, 67450, Pakistan
Email: 1shahidkamal@gu.edu.com, 2adnan.ahmed03@yahoo.com, 3suhail@quest.edu.pk

ABSTRACT

In Web Searching, users are needed to give queries in order to get back the response accordingly. Generally, users get bulk of irrelevant information associated with given input and it needs to be filtered with respect to user requirements about searched contents. Commonly, it is observed that most of queries are unable to describe its purpose, hence known as ambiguous queries. These ambiguous queries establish a noteworthy portion and results in challenging the users' intents towards web search. Therefore, many locations based as well as time based features have been engrossed to deal with such problem in order to achieve information effectiveness in terms of accuracy and relevancy. This paper presents Identifying Ambiguous Queries Method (IAQM) based on Spatial Information, a new method to classify the ambiguous queries based on post search results. In order to originate spatial information from the search results, the ambiguous queries from two different datasets Ambient and Moresque are processed separately by developing a Java-based prototype. The proposed IAQM has achieved improved performance in terms of accuracy as 82% and 78% independently that leads to the motivation of development of small scale search engine in future.

Keywords:Ambiguous, spatial information, disambiguation, information retrieval

1. INTRODUCTION

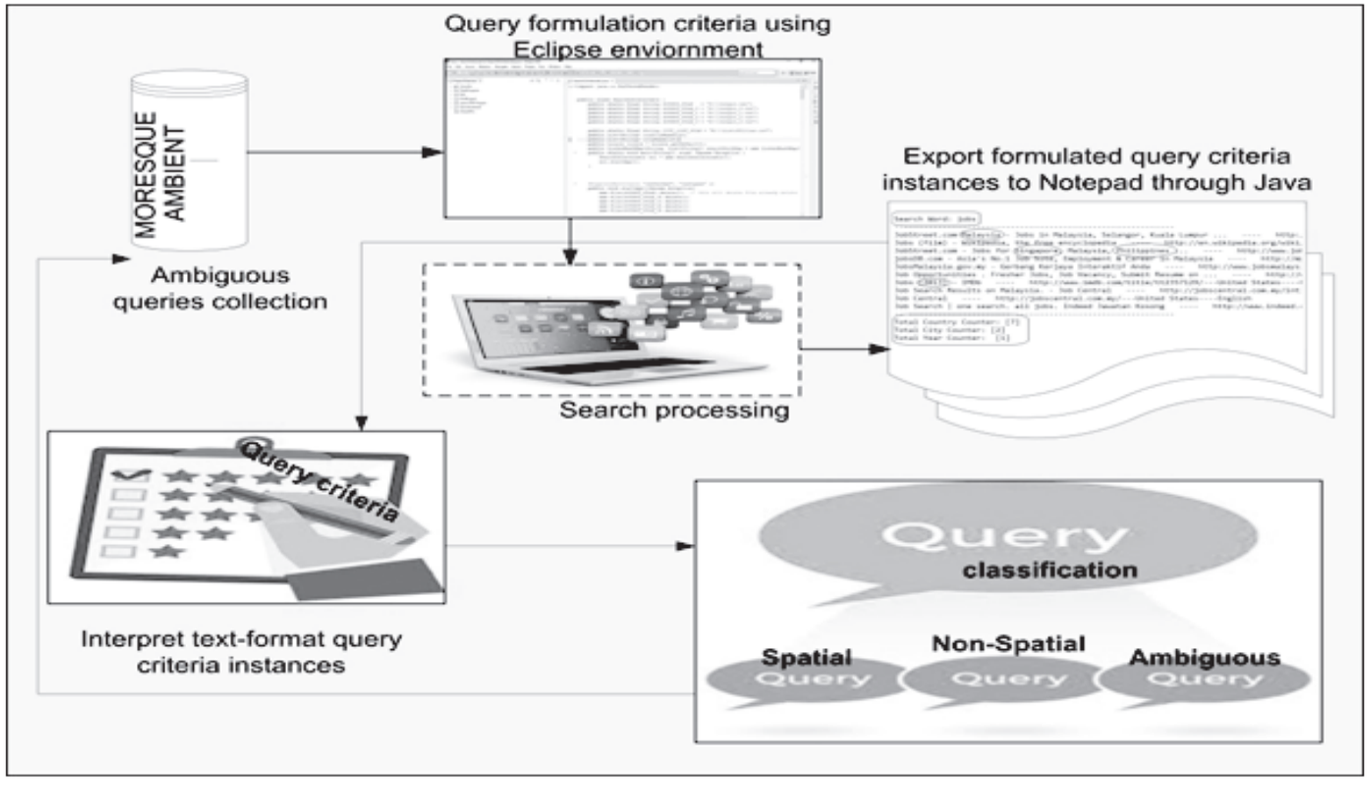
In the field of web search, the essential objective of research is associated with performance enhancement, which is mainly dependent on disambiguation of queries given by the users to find information. For example, when a user gives queries, he/she gets back thousands of results in response of the query which he/she needs to analyze according to requirements. In this process, he/she hold relevant information and disposes off irrelevant one. But however, this process in unviable because time restrictions. Hence it results into finding of the quick way to find relevant information. This situation cause accuracy problem to occur. Lack of domain knowledge leads natural language limitations in the state when users are unable to state their needs effectively [1]. This unclear expression of user requirements creates problem of ambiguity that results in misunderstanding of the queries and their linked results. Such queries are called logically ambiguous and mostly contains short terms i.e., one to three terms only [2]. By locating domain knowledge and initiating refinement process, ambiguity can be resolved by using some features of type spatial or temporal. This process of introducing additional spatial or temporal features leads to spatial search [3-5] along with temporal search [6-8]. Henceforth in this paper, we propose a method named as Identifying Ambiguous Queries Method (IAQM) based on Spatial Information dealing with ambiguous queries to make them clear so as to get accurate search results in response.

The remaining contents of the paper are organized as follow; Section 2 is about description of previous related work. Section 3 elaborates the proposed method and then Section 4 is to give description about. Finally, concluding remarks along with future directions are penned out in Section 5.

2. LITERATURE REVIEW

By using search engines, the purpose of web search is to enhance the process in order to get precise information needed by users. But however, it is becoming challenging because of rapid growth in size and complexity of Internet. To retrieve information, search engines are in need of input queries to be processed upon. These queries are observed as ambiguous[9]; and cause renouncing of performance of search engines in terms of accuracy. Additionally, identification of these ambiguous queries is considered as

thorny task. Hence, the process disambiguation is meant to deal with retrieving relevant information and also identifying the ambiguous queries as well [10]. For this purpose, different disambiguation techniques [8-11] were introduced in past with features of identifying those ambiguous queries. Ricardo, et al. used temporal features in terms of year to process text queries for disambiguation [3]. Clustering approach was utilized to form clusters on the basis of temporal features. However, that approach was lack of accuracy because of solely relying on temporal features as it was argued by many researchers in support of other features like spatial. And also it was contrary to the principle of traditional IR system that doesn't use these features in retrieved documents until they exist in variety [12]. Long ago, in multi-computer applications the concept of spatial information has been introduced by targeting different research areas like user modeling, and information retrieval[13-15]. The spatial information is defined as the information, which is used to illustrate nearby applications with the situation [15]. Then after, the problem of query disambiguation has been researched by [8] utilizing spatial information. Our proposed method IAQM is distinguished with authors in use of spatial information in the form of location being found in results retrieved back in response of user queries. However, our work is differentiated with the authors in terms of using spatial information representing any location in search results produced in response of the queries rather than using separately. Even though, numerous approaches, the determinations need progressing with respect to particular case of using spatial information for query disambiguation.



FigureError! No text of specified style in document.. 1: Work diagram of the proposed method

3. PROPOSED WORK

Our proposed methodology is grounded on hybrid approach and is focused upon primarily on identification of ambiguous queries and then developing a method for disambiguating the contents accordingly. The proposed IAQM is executed to deal with ambiguous queries so as to make them unambiguous such that to retrieve accurate information according to user needs.

In proposed method, we process explicit queries consisting of spatial information in receptive results for refinement procedure. The ambiguous queries are being input from two different datasets; AMBIENT and MORESQUE separately. Furtherance in investigating categories of the queries, we use spatial feature i.e., consisting of information about any place (See Figure 1). In next step, these queries are classified into spatial, non-spatial, or ambiguous based on results with respect to spatial information presence. The spatial category represents information about any place; Non-spatial is representing temporal information in the form of year, while ambiguous ones represent none of

either spatial or temporal information. Thereafter analyzing the receptive contents, queries are being classified and then datasets are updated to be processed in future by other researchers. In order to implement IAQM, the Eclipse IDE on top of a Windows 10 pro (64-bit Operating System) with 4GB RAM and 2.30 GHz Core ™ i5-4200U CPU @ 1.60 GHz are used.

4. RESULTS AND DISCUSSION

In order to assess IAQM, AMBIENT (Ambiguous Entries), [16] and MORESQUE (MORE Sense-tagged QUEries) [17]are used to evaluate its efficacy. The prior dataset comprised of 44 queries and subsequent consists of 114 ambiguous queries. The information about number of queries being processed in our method is presented in the table 1 given below.

Table 1: Information about ambiguous Queries in datasets

Sr.	Dataset	Number of Ambiguous Queries
1	AMBIENT	44
2	MORESQUE	114

Keeping in focus of ambiguity of the queries, our disambiguation approach [18] classify these queries as spatial, non-spatial, and ambiguous. We used search results in order to get executed our method. Upon retrieving 10 search results in form of web snippets, we analyzed one by one to find spatial information such that classification can be made. Figure 2 presents the execution and the analyzing process.

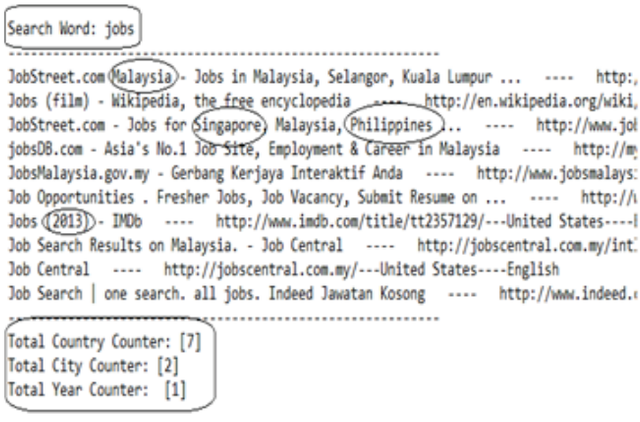


Figure 2: Execution and analysis process

In the above Figure 2, we set "country" and "city" counters for the spatial information so as to target place, while "year" counter variable is used to search out temporal information. The table 2 below presents in first row total number of ambiguous queries being processed. The second row shows the number 23 and 21 as spatial queries based on values retrieved against each variable, 13 and 68 are classified as non-spatial , so as to make total of 36 and 89 clear queries as whole in each dataset. After final results, we come to conclusion that our criteria to classify queries as spatial or non-spatial performed well and hence at last only a small portion of 8 and 25 are left as an ambiguous to be processed in future with the introduction of new features.

Table 2: Results summary after execution of IAQbSIM

Processed queries / DATASET	AMBIENT	MORESQUE
Ambiguous queries	44	114
Spatial queries	23	21
Non-spatial queries	13	68
Total Cleared Queries	36	89
Leftovers ambiguous queries	8	25
Performance achieved	**82%**	**78%**

The figure 3 presents the queries being processed in our proposed IAQM based on the information being presented above in table 2. The blue colored bar shows the values obtained from AMBIENT data set while red colored bar shows values that are obtained after use of queries of MORESQUE dataset.

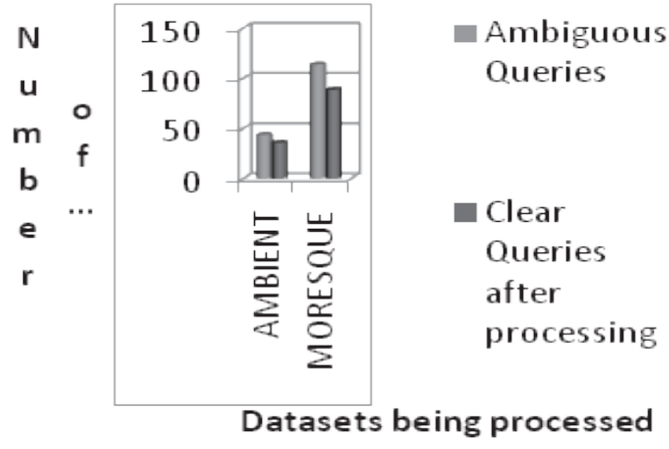


Figure 3: Results after processing datasets

While in figure 4 the query classification information is being shown for the transparent understanding.

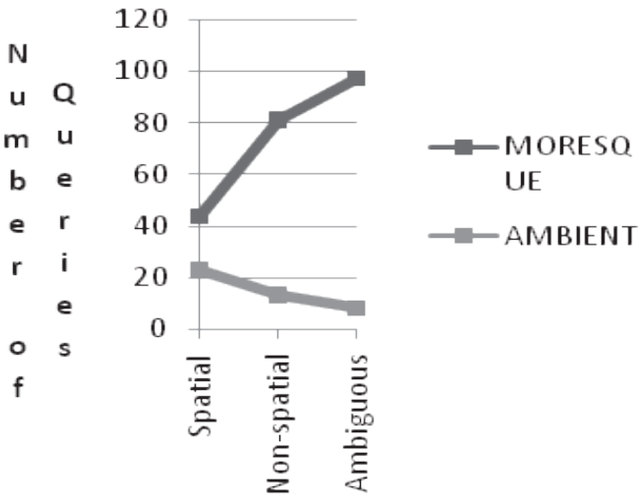


Figure 4: Queries classification chart

The performane of our proposed method IAQM according to defined criteria is depicted in figure 5. The the outcomes of our proposed method IAQM are showed in blue-colored bar while red-colored bar is to show the leftover ambiguous queries, that need to be treated in future.

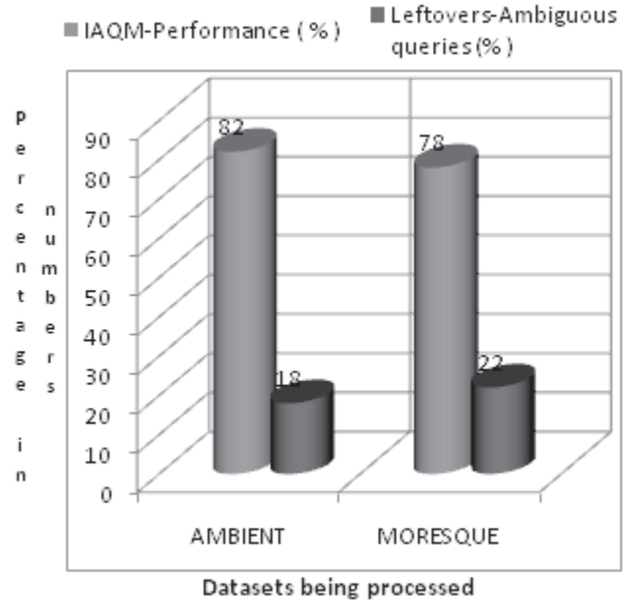


Figure .5: Results come out from different datasets

Furtherance in description of figure 5, IAQM performance and leftover ambiguous queries are shown by two legends respectively in blue-colored and red-colored bars. The X-axis represents the altered datasets that we used for the investigation while Y-axis is demonstrating the percentage that have been symbolized as numbers. Among the results it has been observed that our method IAQM significantly contributed to make the ambiguous queries as clear, while the leftover ambiguous queries i.e., 18% and 22% in the datasets AMBIENT and MORESQUE respectively need to be processed and analyzed by the researchers in future by introducing new features.

5. CONCLUSION

In this paper, we underlined the problems of identifying ambiguous queries input for the purpose of finding relevant information from a majority of the retrieved queries; which is then difficult to examine by users in response of ambiguous queries. In this paper, we recycled a total of 158 ambiguous queries from two different datasets separately with the purpose of making them vibrant so as to enhance the accuracy of the retrieved information. We carried out the process by using spatial features being present in post search results. Our proposed method titled IAQM achieved better performance and filtered 82% and 78% of clear queries in the datasets distinctly, that were known to be ambiguous in past. We also have objectives to test our method by using divergent datasets in order to verify its robustness. In addition, we also have future plan to execute a full-text analysis by combining spatial features with time-based features so as to develop a small scale search engine.

REFERENCES

[1] Zhang, H. and E. Adviser-Jacob, Query enhancement with topic detection and disambiguation for robust retrieval. 2013.

[2] Roul, R.K. and S.K. Sahay, An effective information retrieval for ambiguous query. arXiv preprint arXiv:1204.1406, 2012.

[3] Campos, R., et al. Disambiguating Implicit Temporal Queries by Clustering Top Relevant Dates in Web Snippets. in Web Intelligence and Intelligent Agent Technology (WI-IAT), 2012 IEEE/WIC/ACM International Conferences on. 2012. IEEE.

[4] Drew, T. and J.M. Wolfe, Hybrid search in the temporal domain: Monitoring an RSVP stream for multiple targets held in memory. Journal of Vision, 2012. 12(9): p. 1276-1276.

[5] Lan, R., et al., Temporal search and replace: An interactive tool for the analysis of temporal event sequences. HCIL, University of Maryland, College Park, Maryland, Tech. Rep. HCIL-2013-TBD, 2013.

[6] Kraft, R., et al. Searching with context. in Proceedings of the 15th international conference on World Wide Web. 2006. ACM.

[7] Mizzaro, S. and L. Vassena, A social approach to context-aware retrieval. World Wide Web, 2011. 14(4): p. 377-405.

[8] Anastasiu, D., C., et al., A novel two-box search paradigm for query disambiguation. World Wide Web, 2013. 16(1): p. 1-29.

[9] Song, R., et al., Identification of ambiguous queries in web search. Information Processing & Management, 2009. 45(2): p. 216-229.

[10] Bunescu, R.C. and M. Pasca. Using Encyclopedic Knowledge for Named entity Disambiguation. in EACL. 2006.

[11] Mihalcea, R. and A. Csomai. Wikify!: linking documents to encyclopedic knowledge. in Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. 2007. ACM.

[12] Jones, R. and F. Diaz, Temporal profiles of queries. ACM Transactions on Information Systems (TOIS), 2007. 25(3): p. 14.

[13] Salton, G. and C. Buckley, Term-weighting approaches in automatic text retrieval. Information processing & management, 1988. 24(5): p. 513-523.

[14] Singhal, A., Modern information retrieval: A brief overview. IEEE Data Eng. Bull., 2001. 24(4): p. 35-43.

[15] Tamine-Lechani, L., M. Boughanem, and M. Daoud, Evaluation of contextual information retrieval effectiveness: overview of issues and research. Knowledge and Information Systems, 2010. 24(1): p. 1-34.

[16] C., C. and R. G., Ambient dataset.http://credo.fub.it/ambient/, 2008.

[17] Roberto Navigli and G. Crisafulli. Inducing Word Senses to Improve Web Search Result Clustering. in In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010). 2010. MIT Stata Center, Massachusetts, USA