A Hybrid Query Disambiguation Adaptive Approach for Web Information Retrieval

Roliana Ibrahim[1], Shahid Kamal[1], Imran Ghani[1] and Seung Ryul Jeong[2,*]
[1] Faculty of Computing, Universiti Teknologi Malaysia (UTM), 81310 skudai, Johor Malaysia
[e-mail: roliana@utm.my, skamaltipu@gmail.com, imran@utm.my]
[2] Graduate School of Business IT, Kookmin University, Korea
[e-mail: srjeong@kookmin.ac.kr]
[*]Corresponding author: Seung Ryul Jeong

Received February 23, 2015; revised May 20, 2015; accepted June 23, 2015; published July 31, 2015

Abstract

In web searching, trustable and precise results are greatly affected by the inherent uncertainty in the input queries. Queries submitted to search engines are by nature ambiguous and constitute a significant proportion of the instances given to web search engines. Ambiguous queries pose real challenges for the web search engines due to versatility of information. Temporal based approaches whereas somehow reduce the uncertainty in queries but still lack to provide results according to users aspirations. Web search science has created an interest for the researchers to incorporate contextual information for resolving the uncertainty in search results. In this paper, we propose an Adaptive Disambiguation Approach (ADA) of hybrid nature that makes use of both the temporal and contextual information to improve user experience. The proposed hybrid approach presents the search results to the users based on their location and temporal information. A Java based prototype of the systems is developed and evaluated using standard dataset to determine its efficacy in terms of precision, accuracy, recall, and F1-measure. Supported by experimental results, ADA demonstrates better results along all the axes as compared to temporal based approaches.

Keywords: Disambiguation, Query categorization, Hybrid, Context, location, Information retrieval

http://dx.doi.org/10.3837/tiis.2015.07.008 **ISSN : 1976-7277**

1. Introduction

World Wide Web (WWW) and search engines have get to be vital gears of our everyday life. In spite of huge enhancements being made to optimize the web search over the last decade, still much be done to cope with continually expanding size of the web and needs of the users. Today, web search optimization is an active research area and has gained remarkable attention of experts from both the industry and the academia [1].

One of the significant difficulties in web search lies in inadmissible importance of results caused by ambiguity. Query terms are inherently ambiguous due to polysemy, and most of the queries are short, containing one to three terms only [2]. Consequently the ambiguous queries in terms of user intent and information needs, result into retrieval of many irrelevant pages. As the web size develops, ambiguity gets to be omnipresent and users are in more prominent need for viable method of disambiguation.
The ambiguity can be defined as *"A lack of clear and exact use of words, so that more than meaning is possible*[1]*"*. For instance in Wikipedia[2] when the phrase "World Cup" is searched, it returns 41 entries for different categories including FIFA World Cup, ICC Cricket World Cup, Rugby World Cup, Bandy World Cup, and so on.

In quest of web search optimization, the Temporal Information Retrieval (T-IR) has gained greater attention in recent years [3]. However, majority of these solutions either focus on development of suitable tools or perform behavioral analysis based on log data. Significant numbers of user search queries have strong temporal components or characteristics. These are the queries whose underlying intent may be to obtain newest information, past or anticipated events and largely depend on time. For instance referring to the "World Cup" example, the user might be interested in information about FIFA World Cup 2014 at Brazil. In this regard, if the user issues a query phrase "World Cup 2014", it will make use of the temporal feature and will produce 12 ambiguous results i.e., ICC T20 World Cup at Bangladesh, FIFA World Cup at Brazil, Men's Hockey World Cup at Netherlands, Alpine Skiing World Cup at Austria, FIBA 2014 at Spain and so on. A detailed overview of T-IR, its relevant challenges and opportunities can be found in [4].

Context is an important source of information in computing environments. The term context is defined by the authors of [5] as "any information that can be used to characterize the situation of an entity". An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves. According to the authors of [1], the query disambiguation can be greatly improved by applying contextual information. For instance, referring to our example, if we add the contextual information (Brazil) and rephrase our search query as "World Cup 2014 Brazil", this would produce more accurate results

[1]http://www.macmillandictionary.com/dictionary/british/ambiguity
[2]http://en.wikipedia.org/wiki/World_cup

according to our intent. Hence, it is observable that context plays an important role in resolving the queries ambiguity.

Majority of the existing literature such as rule-based [6], topological [7], and ontological [6, 8] approaches are based on temporal information retrieval. The T-IR based approaches whereas somehow refine the search results by exploiting various temporal features; however, due to lacking of contextual information result into large proportion of irrelevant information retrieval. In this paper, we propose an Adaptive Disambiguation Approach (ADA) that makes use of both the temporal and contextual information thereby retrieving the most accurate results in accordance with the search queries. The proposed approach is comprised of five stages namely: query input, query categorization, sub-query construction, results integration and improved clusters through feedback. Experimental based evaluation of ADA reveals improved performance in terms of accuracy, precision, recall, and F1-measure as compared to existing work.

The rest of the paper is organized as follows: We begin in Section 2 by unfolding the problem background information about the query ambiguity, temporal queries and contextual information used in a query. The related work is depicted in Section 3. Our proposed approach is characterized in Section 4. The implementation of the proposed approach is portrayed in Section 5. The outcomes and assessment is examined in point of interest in the Section 6 though the paper is closed in Section 7 alongside identification of future work.

2. Problem Background

Disambiguating the search intent and improving the accuracy of resulting information is a hot research area where numerous contributions have been made to address these issues. In order to develop a basic understanding of the problem background, we illustrate the compromised accuracy with the help of an example. For instance, **Fig. 1** shows our search query “Cultural Show” given in well-known search engine Google[3] which produced 324,000,000 results in 0.39 seconds. The given term is ambiguous in terms of location and year that’s why resulting so many answers.

[3] http://www.google.com

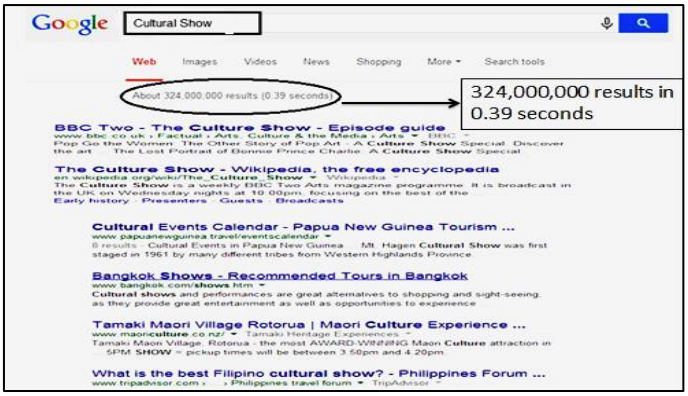

Fig. 1. Results after giving query "Cultural Show"

It is generally accepted that a few queries submitted to search engines are ambiguous by nature e.g. World Cup, Cultural Show etc. Different studies have investigated the inherent ambiguity issues of search queries in various ways. The work presented in [9, 10] describes what are ambiguous queries. For instance referring to word "java" in search query, it is not clear whether the user intent is the application or the Indonesian city. The authors of [6, 7] summarized the following categories of ambiguous queries:

- **Category 1 (Ambiguous Query):** *a query having multiple meanings; e.g. "apple"* which may refer to fruit or company.
- **Category 2 (Broad Query):** *a query that covers a variety of subtopics and a user might look for one of the subtopics issuing another query; e.g. "World Cup 2014"* which covers some subtopics such as FIFA, Hockey, ICC, and FIBA. A user usually issues such a query first and then narrows down to a subtopic.
- **Category 3 (Clear Query):** *a query that covers a narrow topic and has specific meanings; e.g. "FIFA World Cup 2014"*. A clear query usually means a successful search in which a user can find several results with a higher degree of quality in the first results page.

Similarly, a quantification of ambiguous queries is performed in [11] where in GISQC_DS dataset, 220 queries are identified as ambiguous. Likewise, the authors in [12] discuss how common the ambiguity in search queries is. **Table 1** presents a summary of various contributions that have been made for query disambiguation.

Table 1. Existing query disambiguating approaches

Primary Approach	Sub-Approach
▪ Ranking Algorithms	▪ PageRank [8, 9]
	▪ Hyperlink Induced Topic Search [10, 11]
	▪ Weighted Page Rank [12]

▪ Word-Sense Disambiguation	▪ Dictionary and Knowledge-based
	▪ Supervised [13]
	▪ Semi-Supervised or Minimally Supervised[13, 14]
	▪ Unsupervised[14]
▪ Disambiguation and Query Expansion	▪ Generative Concept association model [15]
▪ Clustering	▪ Temporal Clustering through web snippets [16]

Different ranking algorithms are used by search engines (PageRank by Google, HITS by Ask.com) to assess the relevance of results. Similarly, Word-Sense Disambiguation (WSD) approaches deal with the process of identifying the sense of a word in a sentence, especially when the word has multiple meanings.

3. Related Work

In information retrieval, removal of ambiguity is of vital significance both from the user and system perspective [15]. In this section we briefly describe the various contributions that have been made for query disambiguation and search optimization.

Bunescu and Pasca were the first to use Wikipedia as a resource for disambiguation [17]. They expressed the disambiguation task to be a two-step process where a system must first identify the prominent terms in the text and secondly link them accurately. Although 84% of the accuracy has been claimed but it necessitates word connection with time highlights .Bunescu and Pasca's work was initially limited to named entity disambiguation, and then enhanced by Mihalceain [18] thereby developing a more general system that linked all "interesting" terms and achieved 94.33% in terms of precision however keyword word extraction in view of time highlights can be used to enhance the execution.

Ricardo, at al. [16] highlighted the disambiguation of text queries with respect to temporal feature and attained precision, recall and f1-measure as 0.945, 0.92 and 0.943 separately. They proposed a two-stage process where relevant temporal expressions are extracted from results and then grouped into same clusters with respect to common year. Their approach was based on the idea of finding one non trivial term in text and focused on temporal clustering. Alonso et al.[19] first introduced the temporal clustering on the basis of topics and time. Their work solely relies on temporal features thereby compromising the accuracy of the results.

Link Text Topic Model (LTTM) based disambiguation approach has been proposed by Skaggas and Getoor in [20], Although the 61.9% accuracy has been achieved but it resolves the link disambiguation problem only thereby lacking the capability to

disambiguate the user queries. Hence it needed to utilize time highlights in order to enhance the performance.

Boston et al. [15] developed a system (called "Wikimantic") for link disambiguation and query expansion in response to user queries for the retrieval of information graphics. In the developed system, they first disambiguate short text strings, followed by determination of the instant when the sequence of words should be disambiguated. The performance in terms of precision 0.87 has been attained but the main limitation of their system is that it only entertains short queries and the performance is greatly deteriorated when exposed to large queries. Furthermore, it attains low precision and recall as compared to other approaches.

Given a large text string, it's always possible to find at least one trivial term to start the process. Ferragina and Scaiella [21] addressed this problem by employing a voting system that resolved all ambiguous terms simultaneously. The authors made claim of accomplishing 91.7, 89.9, and 90.8 as precision, recall and f1-measure independently. Their system makes use of various characteristics associated with different fragments of the input strings. Furthermore, due to unavailability of trivial terms in short text strings, its performance is greatly affected.

More recently, Anastasiu, D.C, et al [1] investigated the problem of query disambiguation by making use of keywords search and contextual information. First the articles were retrieved on the basis of both combined fragments of query as well as contextual terms. Next they retrieved the articles based on only query terms and finally similarities were computed. The authors accomplished the accuracy over yahoo as 33% and over google as 50% and henceforth can be enhanced by using the term publicizing concept for contextual as well as temporal features. Eventually, the commonly retrieved results were presented to users for their selection. A summary of the related work is presented in **Table 2**.

Table 2. Summary of query disambiguation approaches

Parameters / Article reference	Query types for disambiguation			Dataset Evaluated	Approach					Performance measures				Features		Future work / Limitation
	TX	LK	SR		TF	CS	BN	CL	MO	P	Re	F1	Ac	Tmp	Ctx	
Anastasiu, Gao [1]	√	×	×	Google, yahoo search	√	√	×	×	×	√	×	×	×	×	√	It can be upgraded utilizing customized inquiry in light of time highlights
Boston, Fang [15]	×	√	×	Trec 2007 QA Track	×	×	√	×	×	√	×	×	×	×	√	It can be enhanced utilizing blend of ideas
Campos, Jorge [16]	√	×	×	WC-DS	×	×	√	×	×	√	√	√		√	×	It can made more viable

															by utilization of time with concept
Skaggs and Getoor [20]	×	√	×	Wikipedia 1000 link set	×	×	×	×	√	×	×	×	√	×	√	utilizing time highlights will enhance the execution
Ferragina and Scaiella [21]	√	×	×	IITB	×	×	×	×	×	√	√	√	×	×	√	To overcome vagueness, it needs utilization of term publicizing idea
Alonso, Gertz [19]	×	×	√	Amazon mechanical Turk	×	×	×	√	×	×	×	×	×	√	×	Terms with setting connection can be requisitioned bunching
Bunescu and Pasca [17]	√	×	×	Wikipedia ambiguous queries	×	×	×	×	√	×	×	×	√	×	√	It requires word connection with time highlights
Mihalcea and Csomai [18]	√	×	×	Gold standard	×	√	×	×	×	√	×	×	×	×	√	Keyword word extraction in view of time highlights enhance the execution

Legend:

√	Supported	TF	TF/IDF	PR	Precision
×	Not Supported	CS	Cosine Similarity	RE	Recall
TX	Text	BN	Bayesian Network	F1	F1-measure
LK	Link	CL	Clustering	AC	Accuracy
SR	Search Results	MO	Modeling		
CTX	Contextual Feature		TMP	Temporal Feature	

4. Adaptive Disambiguation Approach (ADA)

In this Section, we give detailed description of our adaptive approach of disambiguating text queries. Basically, we extend the work presented in [22] thereby incorporating the contextual information along with temporal features. Overall our adaptive approach is comprised of five-steps as depicted in **Fig. 2**.

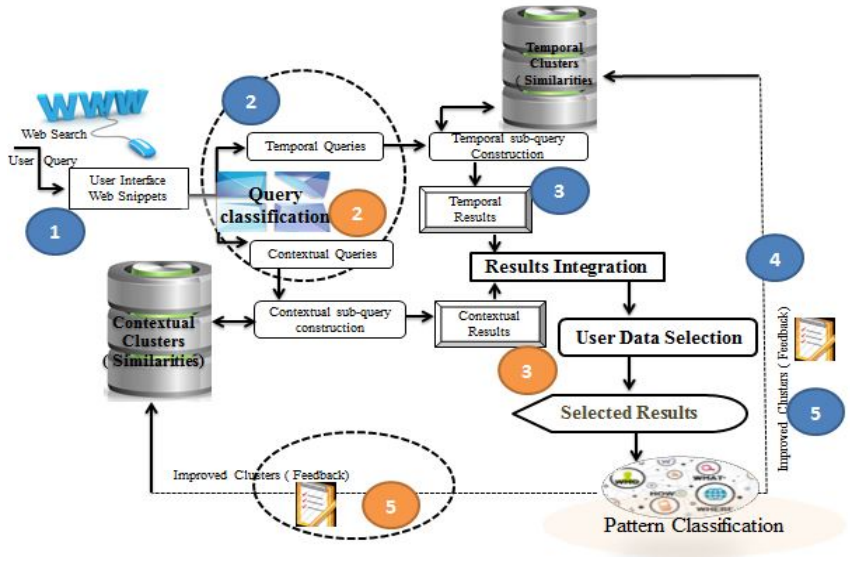

Fig. 2. Block diagram of Adaptive Disambiguation Approach

Furtherance in depiction about **Fig. 2**, there are circles with distinctive hues i.e. blue and orange. The blue shaded circles are intended to demonstrate the procedures while the orange hued circles are speaking to the parallel procedure of same kind utilizing distinctive parameter depicted there as a part of. The five-steps of ADA are given as follows which are further described in the subsequent sub-sections:

- Query Input
- Query Categorization
- Sub-query Construction
- Results Integration
- Improved Feedback

4.1 Query Input

The input query can be explicit (combination of text and location) or implicit (just text). In this paper, we deal with the former one where the availability of contextual information is helpful in refining the search results. We denote the input query by *q* and consider a result set *n* of 10 entries against each query as shown in **Fig. 3**. We use Google Web Services for information retrieval in accordance with the input queries.

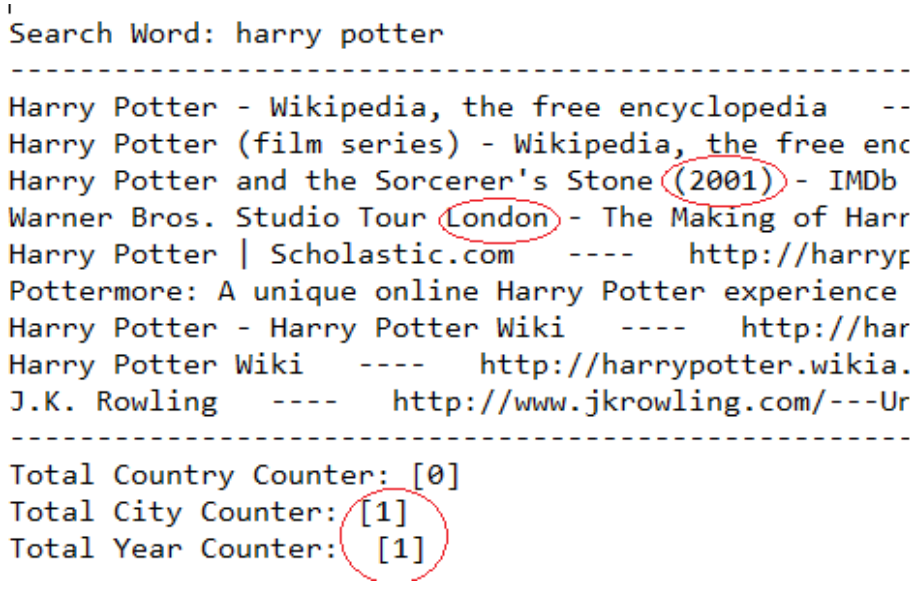

Fig. 3. Search results against in response to query

4.2 Query Categorization

After taking the query from user, next we categorize it in accordance with its concept ambiguity. Similar approach has been adopted in [6, 7, 22] where three different types of concept queries are defined: ambiguous, broad and clear. In our approach, we further classify the ambiguous queries on the basis of both contextual and temporal information into contextual (location based), non-contextual (year based), and ambiguous (neither ensuring contextual nor temporal information) as shown in **Table 3**.

1.2.1 Contextual Queries

These are the queries which are delicate to location i.e. country or city name are recorded in the search results e.g. “poems” in the row 6 of **Table 3** demonstrating the "1" as genuine esteem under the section "Country Information" while displaying the "0" as false under "year" and "city" sections independently.

1.2.2 Non-contextual Queries

The queries are said to be non-contextual if the corresponding results include some temporal features i.e. year e.g. Spartacus, eclipse, harry potter etc. demonstrating the value “1” under the column “year” in the rows 7,8,and 5 of **Table 3**.

1.2.3 Ambiguous Queries

Ambiguous queries are those where no contextual information (country or city) and temporal features (year) are retuned in the search results after query execution e.g. art, books, amazon etc. as shown in rows 1, and 4 of **Table 3** demonstrating "0" value under all columns.

Table 3. Results recorded after execution

Query	Category	Country Information	City	Year	
books	Literature	0	0	0	
amor	Society	0	0	1	
art	Literature	0	0	0	Ambiguous
amazon	Business & Economics	0	0	0	
harry potter	Literature	0	1	1	Non-contextual
poems	Literature	1	0	0	Contextual
spartacus	Entertainment	0	0	1	
eclipse	Other	0	0	1	

As our approach is restricted to only two parameters namely, contextual and temporal information in the search results, hence these remaining ambiguous queries need to be refined in future by defining some other parameters like date of creation, author etc.

4.3 Results Filtering

In this phase, the obtained results are filtered based on the nature of the information found in query i.e. temporal as well as contextual. Based on temporal and contextual evidence, the search results are maintained in two separate databases "DbTemporal" and "DbContext".

4.4 Results Integration and Selection

After organizing the results into corresponding databases, next similarities are computed in order to extract the most desired features from the records in each database. Next, the computed similar results from both the databases are integrated and presented to the users for selection according to their preferences.

4.5 Users' Feedback

In last step of ADA, user will select the provided results according to their requirements. On the basis of user selection implicit feedback will be recorded. The implicit feedback mechanism will help in collecting user preferences to be used for search refinement, and

providing appropriate query expansion suggestions in the future based on their track record.

5. Implementation of the Proposed Approach

Adaptive Disambiguation Approach (ADA) is implemented as middleware between users and Google search engine. For the implementation of ADA, we used the Eclipse IDE on top of a 64bit system with 8GB RAM and 2.5 GHz Corei5 processor. The Eclipse IDE offers the base workspace and an extensible plug-in system for customizing the environment. We implemented the underlined algorithm of ADA using JAVA programming language. The algorithm of ADA is given in **Fig. 4**, which presents our hybrid approach.

```
Algorithm Disambiguation (Query, Location, Year, n)
  [This algorithm returns results containing temporal as well
  as contextual information, where n represents the number of
  search results]
      1. location <- country or city information in search
         result
      2. n <- 1
      3. context <- set of results containing location
         information
      4. Temporal <- set of results containing year
         information
      5. if n < 10 then
      6. for each (location or year) in result do
      7. signature <- set of words in query
      8. result <- MATCH (signature, context, temporal)
      9. location <- country name or city name
      10.    temporal <- year
      11.    end return (result)
      12.    n <- n+1
      13.    end if
      14.    end for
      15.    end algorithm
```

Fig. 4. Disambiguation algorithm of ADA approach

The system, we developed to implement our hybrid ADA approach makes use of six different classes namely: *Menu*, *Disambiguation*, *SearchonInternet*, *SearchEngine*, *Summarize and Feedback*. Initially the Menu class executes a method *getinput()* to take a query for processing (in this case, we are using 220 queries identified as ambiguous in GISQC_DS data set [23], (**see appendix**)). After taking the input, the *Disambiguation* initiates a constructor SearchonInternet*()* of the class *SearchonInternet* which then

executes a method *post()* for class *SearchEngine* to be processed over the web. In response SearchEngine gives back the results to SearchonInternet which then suppress the results and send back to Disambiguation class. Next the Disambiguation class executes four methods namely; *Locationcount()*, *CityCount()*, *CountryCount()* and *TemporalCount()* in order to make a summary of the retrieved results.

After performing the summarization of results, the Disambiguation class executes a Boolean method to check the status, whether they are temporal or contextual. After collecting the results, Menu class sends them to the Feedback class as well as execute a method *Display()* to present them to user. Meanwhile, *recordFeedback()* methods is executed to record the results and then after selection made by user, *store()* method collects all the selections made by the user in order to refine the search results in future.

6. Results and Evaluation

In order to evaluate the capability of ADA in disambiguating the input queries, the performance evaluation was carried out in terms of precision, accuracy, recall, and F1-measure. For input queries, the standard GISQC_DS dataset [23] was considered where the performance of ADA was compared with GTE [24] .The GISQC_DS dataset consists of 450 queries manually extracted from Google Insights for Search. Using GTE, out of 450 input queries, 49% are concluded as ambiguous, 39% clear, and 12% broad. The **Table 4** present a summary of the disambiguation by GTE approach. In this table, we have presented different query types that have been identified in the lietrature as ambiguous, clear and broad respectively. Total number of queries treated in this experiment are 450 given in first row of the **Table 4**. Then there are three columns with headings query type, number of queries and query percentage respectively being identified. By observing the data given in the **Table 4**, 220 queries are being identified as ambiguous presenting 49% of the whole query set. The clear queries are of number 176 out of 450 and hence presenting their share in 39% of the query set. Furthermore, 54 queries contributing in 12% of the query set as broad being identified.

Table 4. Query disambiguation results using GTE

Input: Queries (450)		
Query Type	**Number of Queries**	**Percentage (%)**
Ambiguous	220	49
Clear	176	39
Broad	54	12

As our main concern is to address the ambiguity of input queries, therefore, we focused and processed only the 49% ambiguous queries of GISQC_DS dataset. Unlike GTE which considers the ambiguity solely from temporal perspective, our ADA takes

into consideration both the temporal and contextual aspects of search results. Using ADA, we further categorized 49% of the ambiguous queries of GTE into three categories namely: contextual, non-contextual and ambiguous as shown in **Table 5**. By making use of both the temporal and contextual aspects, ADA further enhances the disambiguation of the input queries thereby adding 62 contextual and 51 non-contextual giving 113 (51%) queries to the clear category. While 107 (49%) of the queries remained as ambiguos to be further investigated in future.

Table 5. Query categorization statistics using ADA

Input: Ambiguous Queries (220)			
Query Type	**Number of Queries**	**Percentage (%)**	
Contextual	62	28	62 + 51 = 113 (51 %)
Non-contextual	51	23	
Ambiguous	107	49	

As a result, ADA increases the number of clear queries from 176 to 289 and reduces the number of ambiguous queries from 220 to 107 as shown in **Table 6**. While the number of broad queries remian same as 54 (12%) in **Table 4**.

Table 6. Query disambiguation results using ADA

Input: Queries (450)		
Query Type	**Number of Queries**	**Percentage (%)**
Ambiguous	107	24
Clear	289	64
Broad	54	12

The **Fig. 5** show the relative performance of ADA against GTE in terms of query categorization. In the Figure, the lagend we have used consists of three structues namely; GTE to whom we have compared our approach in terms of ambiguous queries, our approach ADA and then third one is show the performance acheived by our approach in terms of the contributing percenatge of the ambigous queries. In Terms of ambiguous queries based on data from the **Table 4** and **Table 6** the difference between GTE and our approach is -25% i.e. in our approach the Number of ambigous quereis has been reduced from 49% to 24%. Similary in the case of clear quereis the number of quereis has been increased from 39% to 64% and hence caused an improvement of 25% in making the queries as clear to be processed further. Howevere, in case of third category i.e. broad

queries the GTE and ADA produced the same results i.e. 12% and hence remain on same performance.

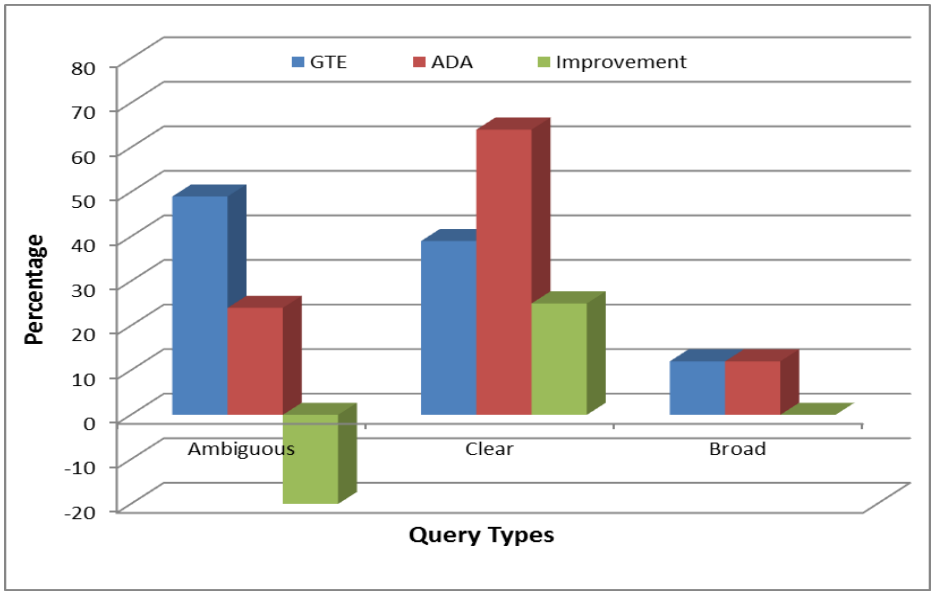

Fig. 5. Performance in terms of queries categorization

The primary goal of query disambiguation is to improve accuracy of the search results according to users aspirations. For the evaluation of an information retrieval mechanism, the most commonly used measures are precision, accuracy, recall, and F1-measure. In information retrieval context, precision (**Eq. 1**), is the percentage relevance of the retrieved documents with the user's information needs. Accuracy (**Eq. 2**), is the degree of closeness of retreived documents to the actual intent of user. Similary, recall (**Eq. 3)**, is the fraction of documents that are relevant to the query that are successfully retreived. Likewise, F1-measure (**Eq. 4**), is the weighted harmonic mean of precision and recall.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{1}$$

$$Accuracy = \frac{\text{TP+TN}}{\text{TP+FP+FN+TN}} \tag{2}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{3}$$

$$F1-\text{measure} = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}} \quad (4)$$

Where True Positive (TP) is the number of locations /years are correctly identified as relevant, True Negative (TN) is the number of locations /years correctly identified as irrelevant or incorrect, False Positive (FP) is the number of locations /years wrongly identified as irrelevant and False Negative (FN) is the number of locations /years wrongly identified as relevant. In the wake of performing the expriment, 77, 79, 45 and 29 rundown search results are recognized as TP, TN, FP and FN independently if there should be an occurrence of GTE assessment while 117, 88, 9 and 6 results are being distinguished as TP, TN, FP and FN individually in assessment of our methodology ADA. Putting these got values in the above mathematical equations (1), (2), (3) and (4) we get the results as given below.

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \gg \frac{117}{117+9} = 0.928 \quad (1)$$

$$Accuracy = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{FN}+\text{TN}} \gg \frac{117+88}{117+9+6+88} = 0.931 \quad (2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \gg \frac{117}{117+6} = 0.951 \quad (3)$$

$$\text{F1}-\text{measure} = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}} \gg \frac{2*0.928*0.951}{0.928+0.951} = 0.939 \quad (4)$$

The undergiven **Table 7** Presents the got consequences of our methodology and the outcomes being accomplished while running the examination over GTE. It is being accepted that as for all these performance measures, the ADA beats the GTE approach.

Table 7. Performance evaluation of ADA and GTE

Parameters / approach	**Precision**	**Accuracy**	**Recall**	**F1-Measure**
GTE	0.631	0.678	0.726	0.675
ADA	**0.928**	**0.931**	**0.951**	**0.939**
Improvement	0.297	0.253	0.225	0.264

The relatively better performance of ADA based on the values being presented in **Table 7** is demonstrated in **Fig. 6** that are mainly attributed to the hybrid approach of making use of both the temporal and contextual features.

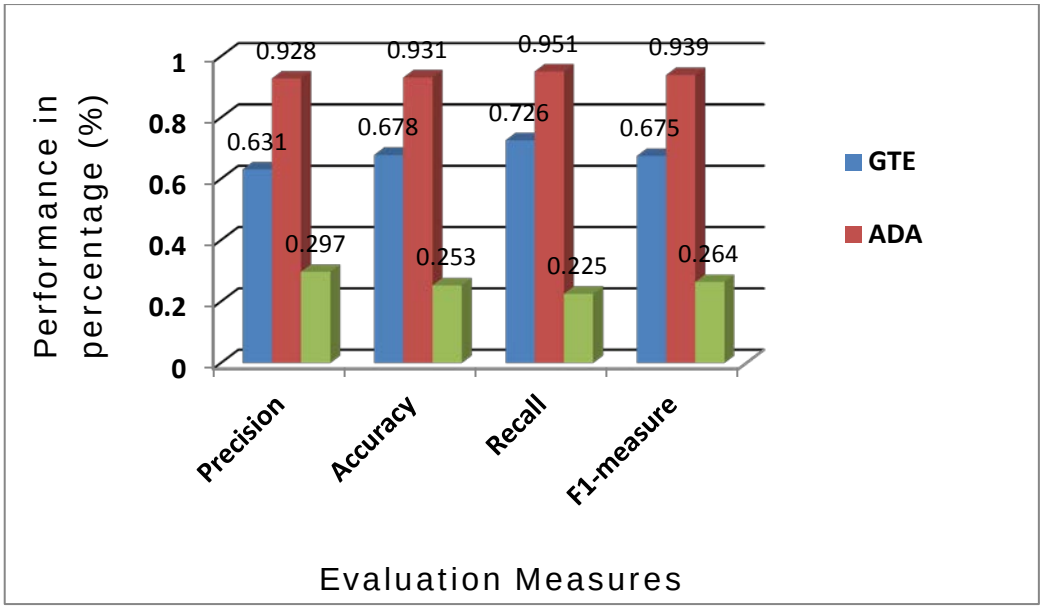

Fig. 6. Performance evaluation in terms of information retrieval measures

7. Conclusion

In this paper, we have proposed Adaptive Disambiguation Approach (ADA) using contextual information gathering of query results, whereas the results are concentrated by location (country/city) and temporal information. Unlike existing approaches, which rely on temporal features in query disambiguation, ADA is based on both temporal and contextual information. By making use of both the temporal features as well as keeping track of user's current location, it addresses the inherent ambiguities in input queries and filters out irrelevant results in response to user's search over the web. Using ADA, a large proportion of ambiguous queries (51% as shown in **Table 5**) are resolved and compared to existing temporal based approaches when tested on standard GISQC_DS dataset. Though 49% of the queries stay ambiguous because of absence of enough data being recovered after execution; thus the remaining queries are held for future work to be recognized in view of some different parameters. Similarly, ADA also outperforms the existing works in terms of the most common information retrieval measures (precision, accuracy, recall, and F1-measure). Part of the future work, we are keen to test ADA using multiple datasets to verify its robustness. In addition, we also plan to develop a small scale search engine which will enable us to carry out a full text analysis using the contextual information in search queries.

Acknowledgement

We might want to say thanks to Universiti Teknologi Malaysia and Ministry of Higher Education (MOHE) Malaysia (Vot No: 4F315) and Research University Grant Scheme (Vot No: Q.J130000.2528.05H84) for the offices and backing to direct this research. In addition we extend our gratitude to Higher Education Commission (HEC) of Pakistan and the Gomal University D.I.Khan.

References

[1] Anastasiu, D.C., et al., "A novel two-box search paradigm for query disambiguation," *World Wide Web*, pp. 1-29, 2013. Article (CrossRef Link)

[2] Jansen, B.J., A. Spink, and T. Saracevic, "Real life, real users, and real needs: a study and analysis of user queries on the web," *Information processing & management*, pp. 207-227, 2000. Article (CrossRef Link)

[3] Joho, H., A. Jatowt, and B. Roi. "A survey of temporal web search experience," in *Proc. of the 22nd international conference on World Wide Web companion*. International World Wide Web Conferences Steering Committee, 2013. Article (CrossRef Link)

[4] Alonso, O., et al., "Temporal Information Retrieval: Challenges and Opportunities," *TWAW*, pp. 1-8, 2011. Article (CrossRef Link)

[5] Dey, A.K., "Understanding and using context," *Personal and ubiquitous computing*, pp. 4-7, 2001. Article (CrossRef Link)

[6] Song, R., et al. "Identifying ambiguous queries in web search," in *Proc. of the 16th international conference on World Wide Web*. ACM, 2007. Article (CrossRef Link)

[7] Song, R., et al., "Identification of ambiguous queries in web search," *Information Processing & Management*, pp. 216-229, 2009. Article (CrossRef Link)

[8] Page, L., et al., "The PageRank citation ranking: Bringing order to the web," 1999. Article (CrossRef Link)

[9] Brin, S. and L. Page, "The anatomy of a large-scale hypertextual Web search engine," *Computer networks and ISDN systems*, pp. 107-117, 1998. Article (CrossRef Link)

[10] Kleinberg, J.M., "Authoritative sources in a hyperlinked environment," *Journal of the ACM (JACM)*, pp. 604-632, 1999. Article (CrossRef Link)

[11] Kleinberg, J.M., "Hubs, authorities, and communities," *ACM Computing Surveys (CSUR)*, pp. 5, 1999. Article (CrossRef Link)

[12] Xing, W. and A. Ghorbani. "Weighted pagerank algorithm," in *Communication Networks and Services Research*, 2004. *Proceedings. Second Annual Conference on*. 2004. IEEE. Article (CrossRef Link)

[13] Stevenson, M. and Y. Wilks, "Word-sense disambiguation," *The Oxford Handbook of Comp. Linguistics*, pp. 249-265, 2003. Article (CrossRef Link)

[14] Schütze, H., "Automatic word sense discrimination," *Computational linguistics*, pp. 97-123, 1998. Article (CrossRef Link)

[15] Boston, C., et al., "Wikimantic: Toward effective disambiguation and expansion of queries," *Data & Knowledge Engineering*, pp. 22-37, 2014. Article (CrossRef Link)

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

Article (CrossRef Link)
[17] Bunescu, R.C. and M. Pasca., "Using Encyclopedic Knowledge for Named entity Disambiguation," in *Proc. of EACL*, 2006. Article (CrossRef Link)
[18] Mihalcea, R. and A. Csomai. "Wikify!: linking documents to encyclopedic knowledge," in *Proc. of the sixteenth ACM conference on Conference on information and knowledge management*, ACM, 2007 Article (CrossRef Link)
[19] Alonso, O., M. Gertz, and R. Baeza-Yates. "Clustering and exploring search results using timeline constructions," in *Proc. of the 18th ACM conference on Information and knowledge management*, ACM, 2009. Article (CrossRef Link)
[20] Skaggs, B. and L. Getoor, "Topic Modeling for Wikipedia Link Disambiguation," *ACM Transactions on Information Systems (TOIS)*, pp. 10, 2014. Article (CrossRef Link)
[21] Ferragina, P. and U. Scaiella. "Tagme: on-the-fly annotation of short text fragments (by wikipedia entities)," in *Proc. of the 19th ACM international conference on Information and knowledge management*, ACM, 2010. Article (CrossRef Link)
[22] Campos, R., A. Jorge, and G. Dias. "Using web snippets and query-logs to measure implicit temporal intents in queries," *SIGIR 2011 Workshop on Query Representation and Understanding*, 2011. Article (CrossRef Link)
[23] Campos, R., "*Google Insights for Search Query Classification dataset (GISQC_DS)*," 2011. Article (CrossRef Link)
[24] Campos, R., et al., GTE-Cluster: "A Temporal Search Interface for Implicit Temporal Queries," *Advances in Information Retrieval*, Springer. pp. 775-779, 2014.
Article (CrossRef Link)

Appendix:

Ambiguous Query Collection (220 Ambiguous Queries)

Dataset (CrossRef Link)

Books	health	map	moon	wm
Amor	cancer	bbc	big bang	nokia
Art	dr	wetter	dna	mobile
Amazon	heart	friv	cool math	blackberry
harry potter	diabetes	backpage	physics	orange
Poems	nhs	bai	big bang theory	samsung
Spartacus	caf	news	fizy	sms
Eclipse	et	cnn	volcan	android
robin hood	dr oz	haber	volcano	htc
Car	sexuality	fanatik	cool maths	apple
Auto	plan b	as	shoes	hotel
Ford	furniture	svz	nike	hotels
Honda	garden	tre	walmart	sncf
Bmw	kitchen	jagran	adidas	travel
Toyota	lowes	noticias 24	allegro	holidays
Cars	farmville	canon	argos	las vegas
Nissan	mattress	camera	oovoo	new york

Mercedes	waka waka	nikon	mercadolibre	pnr status
Chrome	kayu	tagged	deichmann	irctc
Dacia	nook	dslr	etsy	airasia
Face	home	rent	hi5	nba
Hair	ups	real	blog	wm 2010
Tattoo	farm	real estate	chat	memurlar net
Fitness	bp	property	picnik	ssk hizmet dökümü
Spa	poste	rightmove	hyves	fac
Tattoos	free farm	camping	test	prouni
Yoga	ovi	bike	visa	ösym
Avon	u haul	horse	detran	fb
piercing	mail	marathon	act	golf
ipl 2010	msn	patterns	science	sport
ipl	messenger	rc	meb	sports
business	sign in	new kids	cesgranrio	cricket
marketing	web	catfish	cake	vivo
calendar	yonkis	sperry	chocolate	caixa
security	baby	maps	money	video
iş	habbo	traductor	bon coin	me
staples	barbie	translate	le bon coin	mp3
security essentials	names	dictionary	pizza	radio
download	bebe	übersetzer	food	avatar
windows	craigslist	math	game	games
hp	meteo	mario	zing	juegos
firefox	jobs	wow	beer	masterchef
chase	mac	lottery	you	4399
euro	adobe	7k7k	free	pacman
ing	finance			

Roliana Ibrahim is a senior lecturer at the Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia. She has been servicing UTM for more than 10 years after a few years' experience working as a system developer in the industry. She received her BSc (Hons) Computer Studies from Liverpool John Moores University, MSc Computer Science from Universiti Teknologi Malaysia and PhD in the field of Systems Engineering from Loughborough University. Her current research project relates to the development of ontology based data warehousing and data mining model for oral cancer research data repository and improvement of data mining techniques for risk and survival analysis of cancer patients.

Shahid Kamal is a PhD student in the Software Engineering Research Group (SERG) at Faculty of Computing, Universiti Teknologi Malaysia. He received his BSc Computer Science from Gomal University, MSc Computer Science from ICIT, Gomal University D.I.Khan Pakistan. His research includes information systems, data mining, web search and its related issues and information integration. Besides, he is faculty member of the ICIT Gomal University D.I.Khan, Pakistan since 2007.

Imran Ghani is a Senior Lecturer at Faculty of Computing, Universiti Teknologi Malaysia (UTM), Johor Campus. He received his Master of Information Technology Degree from UAAR (Pakistan), M.Sc. Computer Science from UTM (Malaysia) and Ph.D. from Kookmin University (South Korea). He is the member of Software Engineering Research Group (SERG). He teaches Software Architecture and Design, Software Engineering, Secure Software Development and (Software) Application Development. Dr. Imran Ghani has 50+ publications in Journals, Proceedings, and Book Chapters.

Seung Ryul Jeong is a Professor in the Graduate School of Business IT at Kookmin University, Korea. He holds a B.A. in Economics from Sogang University, Korea, an M.S. in MIS from University of Wisconsin, and a Ph.D. in MIS from the University of South Carolina, U.S.A. Dr. Jeong has published extensively in the information systems field, with over 60 publications in refereed journals like Journal of MIS, Communications of the ACM, Information and Management, Journal of Systems and Software, among others. Dr. Jeong's areas of interest are Process Management, Software Engineering, Systems Implementation, and Information Resource Management.