International Journal of Research in Advent Technology, Special Issue, March 2019
E-ISSN: 2321-9637
Available online at www.ijrat.org

Integrating Contextual and Semantic Relevancy in Information Retrieval

S. Shanmugapriyaa
Independent Researcher
Pondicherry, India
shamisealan@gmail.com

***Abstract*—Information retrieval plays an important role in gathering relevant content across resources on the web. The ambiguity of the content in the stored resources and incomprehensibility of the user's need are the major issues in retrieving information. This paper focuses on relevant information retrieval pertaining to user and content aspects. It concentrates on enhancing the retrieved results based on semantic relevance of the sources in the repositories and contextual relevance based on the perspective of the users. The stored repositories are mapped to index based on lexical and conceptual semantics. When a search is performed, all relevant information semantically mapped to the search term are retrieved. The user's preference is analyzed and the current context is derived. Based on this context, the final set of relevant information are presented to the user. This eventually helps in overcoming ambiguity, diversion of search and time constraint.**

Keywords—Information retrieval; contextual retrieval; semantic retrieval; contextual relevancy; semantic relevancy, user context; semantic mapping.

I. INTRODUCTION

Web search process is a common task today. The aim of that task is to get some useful information. Users utilize web search to navigate to a particular website, fetch information regarding their need and to do some transactions. Out of the three, fetching information from the web is very difficult. This is due to the vast resources available on the web. Every resource may not be accurate or clear in all aspects. The process of gathering relevant information from these resources is the main task of retrieval. Precisely, information retrieval is the process of fetching information from stored resources. Information is retrieved either by a human or by a machine based on the scenario. Information retrieved may seem often relevant and sometimes unrelated to the user. Once the results are retrieved, they are ranked as per relevancy and popularity factors. Relevancy is when the retrieved information is partially or fully related to the search term. This is in contrast with the database systems where the results are either exactly matched or partially matched with the search term instead of relevancy criteria [1]. There is more scope for the users in information retrieval to expand and extend their search based on the retrieved results.

There are 2 kinds of users: (1) Low-end users - they merely know to search using a query. (2) High-end users - they exactly know how to proceed for completing a search technically. They ultimately get satisfactory results either by means of search iterations or by using advanced search options. A retrieval system should basically help both these users to get their requirement irrespective of the manner they perform the search. In the usual information retrieval process, users are somewhere forced to change the way they query to the system. All searches do not help users to find out what they want. Instead, some retrieved results divert the users from the context of their search without understanding their need. Thus, there are typically 2 issues to be addressed in this paper namely context and semantics.

Context is the scenario or the circumstance of the user who is in need of information. The requirement always need not be in general but sometimes specific. The system needs to understand in what aspect the user is looking for the information. By gathering user related data, it is possible to enhance contextual relevancy. This relevancy will waive away certain unwanted results which are not in par with the user's circumstance. This will ultimately reduce ambiguity. There are three contexts before the retrieval process: user context, query context and search context. Query and search contexts depend on the term of search without considering the user's need. So, more importance is given to the user context where the user enjoys the ultimate satisfaction with the retrieved results.

Semantics says about the meaning of a term. Lexical semantics talks about grammar and syntax related meaning. Conceptual semantics is all about knowledge domain classification in a broader aspect. Information about a particular topic may be scattered everywhere. Though we can't merge them all, we can still classify them and group them as clusters based on concepts. When a search in initiated,

until the user's context is known, a good retrieval system should fetch information from all related concepts. This becomes the predominant task of every search engine. When all related documents are not collected, the retrieval process is not complete in all aspects.

A. Overview of Information Systems

Information systems work in the order of information seeking, information searching, information retrieval [9], information extraction and information presentation stages. They are bundled inside the generic architecture as shown in Fig.1.

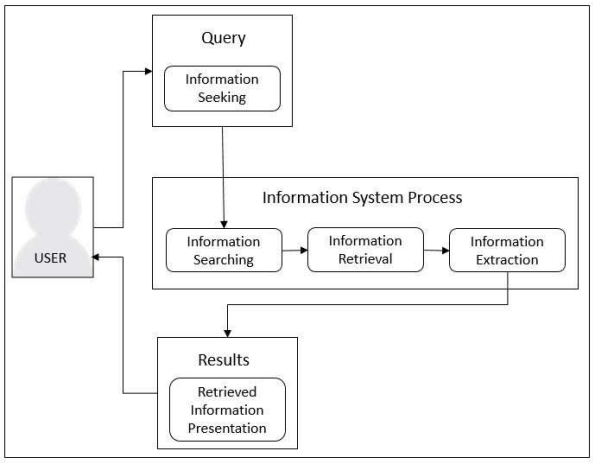

Fig. 1. Overview of Information Systems

The architecture of the system contain four segments. The user segment, query segment, information system process segment and the results segment to be presented to the user. This architecture can work in a loop until the user gets a satisfactory result by making alterations to the query. The user gives the query in the seeking stage which is then passed on to the search stage where actual retrieval work starts. The retrieval stage uses the query and searches for the relevant documents and produces a retrieved relevant set which is further sent to the extraction stage [3] where results are ranked and processes like filtering and summarization may take place. It passes on to the final stage of presentation which is actually viewed by the user. If results have to be still more precise and clear, the user should be able to make changes to the query and the whole process gets repeated until satisfactory levels are met by the user.

B. Importance of Context and Semantics in Retrieval

All users do not have the same circumstance of need of information. Every user has a unique behavior during search process. The query may be the same but the need might be different. Such needs can be studied from the user's previous and current behavior. In case of an online content, the user may hold a personalized profile of his/her needs or interests. Apart from that, browser collects data from the user during his/her visit to a website. All these together when properly assembled may give rise to the context of the user [10] who is querying. But in the present scenario, user context is studied only means of interaction process and mobile-based profiles. Sometimes user feedback is also collected. These actions somewhere indulge subjectivity. The need to study the user behavior more deeply is stressed to know about a user by the system itself rather than manually collecting details from users themselves.

Regarding semantics, there are different meanings for terms and those meanings tend to change when the concept is changed. The words accompanying the query term in the document may sit in different places and in each positon its meaning may differ. Semantic mapping to the terms of document in regard to lexical and conceptual basis will pave way for the terms to be indexed as per the intended meaning occurring in the document position. Lexical semantics include synonyms, hyponyms, hypernyms and homonyms. Conceptual semantics include clustering and classifications of terms according to the content analysis process.

A good search system should not be biased to retrieve results to the user. When the retrieved result is semantically organized in the broader aspect, it is further narrowed down when user's context is applied to it. This helps in inclusion of related documents to the user's query when its semantic dimensions are mapped as well as removing certain non-relevant documents apart from the user's context.

The rest of the paper deals with how to improve relevancy in terms based on the retrieval of related concepts of the search term and also cater to the needs of the users.

II. RELATED WORKS

The works carried out in the field of information retrieval served as a base to think about inclusion and expansion in the existing framework of the retrieval process.

- The work in [2] discussed the design of retrieval system along with the variety of searches performed in it. It gave an idea on how relevant results are retrieved after a search process. A simple user model and query expansion idea was also initiated.
- The article in [4] showed the process of retrieval for a bibliography related information with components like representation of sources and how formal queries are constructed from user queries with certain rules.

- The idea of basic mining algorithms based on keyword, pattern, and sample was discussed in [6] where semantic content of a document was given importance in the indexing process.
- The review paper in [7] considered the formulation of query based on the information need along with overall discussion of retrieval techniques.
- The author in [8] conveyed reference based and multilingual query expansion approaches in retrieval. He also stated the importance of context vectors and categorization using the search interface.
- The survey paper in [13] gave considerable importance to query processing and information extraction. Its future work stated that efficient retrieval is possible by integrating a domain knowledge base to the system.
- The framework proposed in [14] indicated the approach of semantic crawling of the documents based on ontology using WordNet.
- The study of whitepaper in [16] exposed the importance of concept related searching mechanism which ultimately has to depend on conceptual semantics.
- Based on intellectual and theoretical perspectives in [17], critical constructs were formulated like relevance, presentation, similarity between documents, query, interaction, uncertainty principle and neutrality of technology which eventually became important factors in the retrieval process.
- The features of lexical, syntax, semantics, query expansion and search classifiers are pinpointed in [18] with respect to educational web search. The experiment resulted in better relevancy when query classifier was applied.
- The paper in [20] described the interpretation and expansion of queries based on personalization of context along with semantic entities.
- The authors in [21] pointed out the difficulties of search engine users through assessment of the results retrieved for their queries by means of various factors related to usage of web search engines.
- A comparison study of information tools done in [22] reflected that users (post-graduate students) go to more than one channel for retrieving relevant information for their need since they are not able to get all details in a single window. Even when they are provided with ample information, they are not able to select the accurate one and are often stuck with ambiguity.
- The report in [24] mentioned various challenges in information retrieval namely users and context sensitive retrieval, multilingual and multimedia issues, improving objective evaluation and formal models.
- The report in [25] discussed that mere ranking of retrieved result list will not help users. The proposal indicated that the context of the user has to be captured and domain mapping should be introduced along with evaluation metrics.
- The report in [26] projected the need of fairness, accountability, confidentiality and transparency in retrieval and cognitive based user model issues are also discussed.

All these works correlate to state that queries have to be reconsidered before submission to the search system and the results retrieved for the query should not confuse the users with ambiguousness. Thus, the query and ambiguity issues are taken up for study and a model is proposed imposing a way to minimize the above issues.

III. PROPOSED MODEL FOR INTEGRATING CONTEXTUAL AND SEMANTIC RELEVANCY

The model describes inclusion of semantic mapping for content relevancy and modification in user behavior analysis for context derivation.

A. Content Relevancy

Considering the information storage section, there is a repository of documents stored. Primarily, the index for terms occurring in the documents are created using the currently available indexing techniques. This repository is called the index references for terms. This alone will not help in fetching all related documents. Therefore, a combination of related lexical semantics and a table of conceptual semantics are already made available internally in the system. Here, each term in the index get mapped to one or more semantic items from the available list or dictionary by using an efficient mapping technique. This technique can be a data mining tool to classify the words based on clusters and thus terms can be grouped as conceptual clusters. Fig. 2, encapsulates the framework of overall process flow inside the storage section of the system.

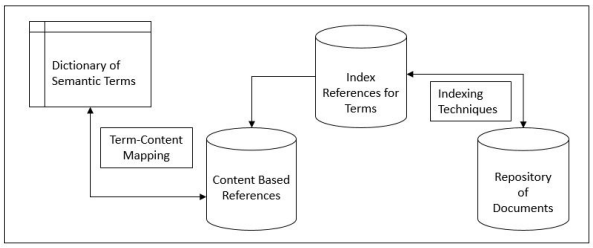

Fig. 2. Framework of Semantic Analysis

The semantic based content has to be readily available whenever a source is added to the repository so that when a search is initiated, access to the information is readily possible without any internal mapping done during retrieval process. To make this possible, a dictionary of lexical semantics and conceptual semantics are connected to the repositories, where mapping is done internally for all sources irrespective of being accessed or not. When this mapping is over, along with index references created for document terms, special references on semantic terms for every indexed term is also created. This also helps users finding content related to their context since context might also be in turn connected to one of the concept based topics.

B. User Behavior

The next section that needs a revision is the user behavior. Whenever the user keys in a query, it is passed on to the search system as a single parameter. When the query is sent, another parameter called the context of the user should also be sent. This parameter can either be hidden from the user in order to prevent misconception or can be explicitly made to be viewed in order to increase user feedback and properly evaluate the system.

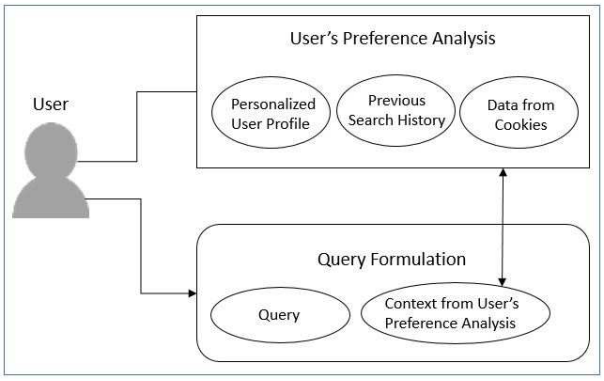

Fig. 3. User Behavior Analysis

An analysis on the user preference has to be done for both the low-end and high-end users. Fig. 3, explains user behavior analysis. The user preference can be studied from three aspects namely the user's personalized profile available both offline and online, the user's previous search history available both online and offline, most of the times collected from browser history and the user's data collected via cookies. These cookies differ from website to website but additionally are also accessed by the latest browsers in studying the user behavior. Using the above three aspects, specific techniques have to be discussed and decided to retrieve the overall context of the user providing the query. This context may also map to one of the concepts already readily available in the storage section of the system as content based reference repository. Thus, the parameters from the user's side are the query itself and the context analyzed using the user preferences. This combination is mentioned as Query Formulation section [5] in this segment. Frequent check has to be made in the user profile to keep it updated every time a user keys in to start to search and also after a search is completed.

Some of the user data include language, interests, education level, background of the search area, location, access to a particular device and any disability on the user's side. The gathered user data to derive the context stays in the client side. An anonymized profile is sent to the server for retrieval just with the derived context. This prevents leakage of user data thereby enhancing user data privacy.

C. Search and Retrieval

In this segment, we can see how a relevant result set is being retrieved from the resource repository. Fig. 4, consolidates the exact process happening in the retrieval of information. It has to be looked up from the bottom. There are internally three sections called search, retrieval and storage. The storage section is ready with its resources, index and content based references after being semantically mapped. Search section provides the query parameter to the system and the retrieval system searches for it in the index collection. Once found relevant, those documents are retrieved by the system. This retrieved set is checked for semantic relevancy against the content based references already available on the storage section. The search is now broadened to the content based retrieved result set [19]. Once all semantically relevant documents are retrieved, it passes on to the next stage of matching the context. This context is derived from the query formulation section of the user behavior segment. It is also mapped against the content based references which also has conceptual references inside it. A technique used to study and analyze the context in terms of concept is intrigued internally before the final set of result is retrieved. Once the mapping is done, we get the final retrieved result set which is now contextually and semantically relevant. This set of to and fro operations keep changing internally in the retrieval system whenever the user preference is modified and there is a change in the accessed document repository.

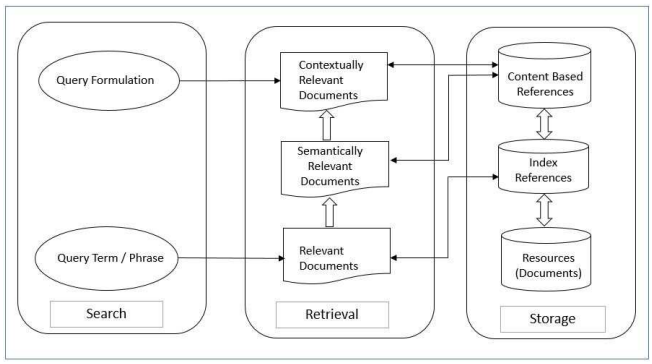

Fig. 4. Architecture of Information Retrieval System

The output of the information retrieval process becomes the input to information extraction process. Only when reliable and relevant information is retrieved at this stage, it can be sent to further process. The extraction segment which covers extracting techniques from the so far retrieved results is beyond the scope of this paper. So, the detailed process of the segment is not discussed here.

When the results are presented to the user, the broad context of the retrieval based on the user context should also be made known to the user which helps the user to connect with the result easily and comprehensibly without doubts.

D. Pseudo-Procedure of the Model:

The procedure is split into three parts namely server side, client side, retrieval and results.

1) Server side:

a. There are a collection of documents in the web.
b. Every document is individually crawled by spiders and an index based reference is generated for every term in the document.
c. A dictionary of lexical semantics and domain base hierarchy are already internally stored.
d. Content analysis is done on the documents and they are categorized into different concepts based on the occurrence of individual term and phrasal terms, thus giving rise to conceptual semantics.
e. A mapping is done among conceptual semantics, lexical semantics and the index terms thereby creating a reference database with each term or term phrase and its related semantics.

2) Client side:

a. User's search history, personal profile from online and offline accounts and user data collected from cookies are analysed.
b. User's likes, interests, preference data are collected.
c. User's activity is also monitored and updated.
d. Using the collected data, the genre of search is decided and ready as a parameter in the user side as context.
e. The user keys in the query.
f. The query along with the context is sent into the search system.
g. The context is sent as an anonymized profile to protect privacy.

3) Retrieval and Results:

a. The query term is first sent for a search.
b. If a term is looked upon, it's related lexical semantics namely synonyms, hyponyms, hypernyms, homonyms are all looked upon.
c. Similarly, the concepts related to the term are also fetched.
d. All fetched results from b and c are retrieved and temporarily stored in the memory.
e. This set in d is the semantically relevant result set.
f. The context of the user is now supplied and gets the conceptual reference as in c.
g. From the result set e, the context – concept reference is searched.
h. It gives a new result set which is relevant to the user context.

E. Illustration

The user keys in the query – 'Oreo'. When this term is searched on the web, the general Oreo search lands in the results with the term occurrence 'Oreo' in the documents. The usual retrieval of documents can be either related to cookies and biscuits because of its common usage or related to all documents having the term irrespective of its kind.

The proposed model will retrieve a temporary result set with all kinds of concepts related to 'Oreo' (semantic relevance) and the final result set will retrieve based on user's need (contextual reference). This difference in the final result set retrieved is illustrated in Fig. 5. The result set of usual retrieval will fetch Doc1 through Doc12 just based on the term occurrence, whereas only Doc1 through Doc7 are semantically related to the term. From this temporary set, based on context of the user, Doc1 through Doc3 are retrieved.

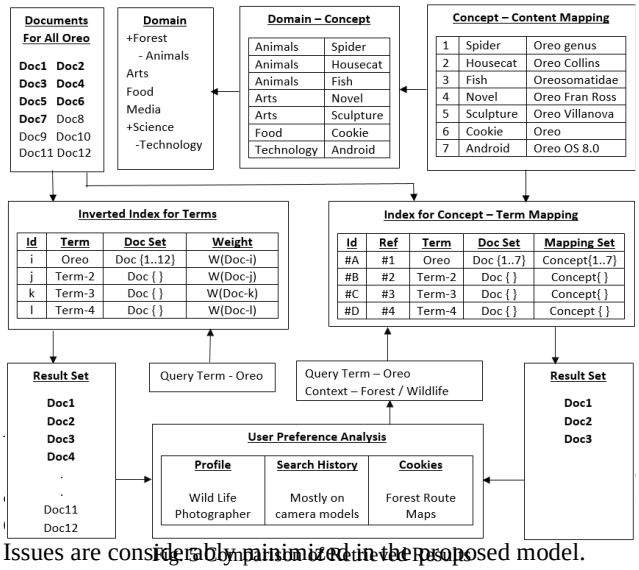

Fig. 5. Comparison of Retrieval Results

Issues are considerably minimized in the proposed model.

TABLE I. COMPARISON OF APPROACH TOWARDS ISSUES

No	*Issue*	*Existing Retrieval*	*Proposed Retrieval*
1.	Fairness	Possibility to be unfair to a group of users.	Possible since terms in content are mapped to concepts.
2.	Accountability	Retrieved set is ranked by popularity factor which is assumed by other users' actions.	All documents are mapped to a concept based index which confirms relatedness avoiding guesswork.
3.	Confidentiality	User's personalized profile is shared to server. Possibility for leakage of data.	User's personalized profile is at client side and only context is shared to server by anonymity.
4.	Transparency	Possible only for location of resources.	Semantic mapping shows the route to retrieved results.
5.	Search engine Bias	Possibility of resource based or user based bias.	User based bias reduced due to individual preference analysis.
6.	Relevance	Presence of term with maximum term weight irrespective of context.	Semantically related to the search term and context of the user.
7.	Context	Query or search context.	User context.
8.	Semantics	More lexical and less conceptual	Lexical and conceptual on equal scales.

IV. APPLICATIONS

The modified sections of the system when included in the present retrieval system, will pave ways to get more relevant results. It will also remove certain unwanted results which are just included merely for the occurrence of a term and repeated in a document irrespective of its actual need. The effects can be visualized in the ways as follows:

- When this modification is applied on the web search engines, this will stop search engine optimization providers to spam with abundant and redundant keywords in their websites which induces irrelevant clicks and social media activity from the public.
- It also plays a vital role in the ranking algorithm of search engines to stop looking only at the keywords and rank the websites based on the original content rather than going behind inappropriate keyword jamming.

V. CONCLUSION AND FUTURE WORKS

The contents of the documents are mapped both lexically and conceptually. This semantic mapping will gather all relevant documents related to the search term. The context of the user is included in the query so that the user's exact need will be known during the search process. This context has to be applied only when the relevant documents are fetched after semantic mapping. Only then the results will be scrutinized as per user's need. Applying context before semantics might miss some relevant documents since context restricts down the search process to include fewer results instead of not broadly expanding its approach. This optimizes the search-retrieval time since users can minimize the iterations of search to get satisfied results. This model also paves way to be fair, accountable, keep user profile confidential and considerably transparent compared to the existing scenario.

In the future, works can be extended by presenting new techniques to enhance the content based mapping. Similar enhancement can also be done by applying efficient personalization techniques in order to study the user behavior.

REFERENCES

[1] O. Yilmazel, B. Yurekli, B. Yilmazel and A. Arslan, "Relational databases versus information retrieval systems: A case study", Proceedings of the IADIS International Conference Applied Computing, Italy, pp. 273-276 , November 2009.

[2] IGNOU, "Unit-3 Information Retrieval Systems", 2019 [Online]. Available at: http://egyankosh.ac.in//handle/123456789/25567.

[3] A.M. Robertson and R. Gaizauskas, "On the Marriage of Information Retrieval and Information Extraction", Proceedings of the 19th Annual BCS-IRSG Colloquium on IR Research, Aberdeen, Scotland, J. Furner and D. Harper, Eds. London: Springer-Verlag, April 1997.

[4] M. K. Buckland and C. Plaunt, "On the Construction of Selection Systems", School of Library and Information Studies, University of California, Berkeley, pp: 15-28, 1994 [Online]. Available: http://people.ischool.berkeley.edu/~buckland/papers/analysis/analysis.html.

[5] A Singhal, "Modern Information Retrieval: A Brief Overview", Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 2001; 24: 35–43.

[6] J. Lasic-Lazic, S. Seljan and H. Stancic, "Information Retrieval Techniques", Croatia: Proceeding of 2nd CARNet Users Conference, 2000.

[7] A. Roshdi and A. Roohparvar, "Review: Information Retrieval Techniques and Applications", International Journal of Computer Networks and Communications Security, vol. 3, no. 9, pp. 373–377, September 2015.

[8] Charles H. Heenan, A Review of Academic Research on Information Retrieval, 2002 [Online]. Available at: http://eil.stanford.edu/publications/charles_heenan/AcademicInfoRetrievalResearch.pdf.

[9] D. Inkpen, "Information Retrieval on the Internet", Volume III, Part 3, 213, University of Toronto, Canada, 2006 [Online]. Available at: http://www.site.uottawa.ca/~diana/csi4107/IR_draft.pdf

[10] Igor Jurisica, "How to retrieve relevant information?", Proceedings of the AAAI Fall Symposium Series on Relevance, New Orleans, Louisiana, 1994, pp. 101–104.

[11] M. Sahami, V. Mittal, S. Baluja and H. Rowley, "The Happy Searcher: Challenges in Web Information Retrieval", In: Zhang C., W. Guesgen H., Yeap WK. (eds) PRICAI 2004: Trends in Artificial Intelligence, pp: 3-12. Lecture Notes in Computer Science, vol 3157. Springer, Berlin, Heidelberg.

[12] D. Lewandowski, "Web searching, search engines and Information Retrieval", Information Services & Use, vol. 25, no. 3-4, pp. 137-147, 2005.

[13] R. Sagayam, S.Srinivasan and S. Roshni, "A Survey of Text Mining: Retrieval, Extraction and Indexing Techniques", International Journal Of Computational Engineering Research (ijceronline.com), Vol. 2, Issue. 5, pp.1443-1446, September 2012.

[14] H.M. Harb and K.M. Fouad, "Semantic Retrieval Approach for Web Documents", International Journal of Advanced Computer Science and Applications, Vol. 2, No. 9, pp. 67-76, 2011.

[15] K. P. Bachchhav, "Information Retrieval: search process, techniques and strategies", International Journal of Next Generation Library and Technologies, Vol. 2, Issue 1, pp.1-10 , February 2016.

[16] J. Challis, "Lateral Thinking in Information Retrieval White Paper", Information Management and Technology, Vol. 36, Part 4, August 2003.

[17] B.J. Jansen and S.Y. Rieh, "The Seventeen Theoretical Constructs of Information Searching and Information Retrieval", Journal of the Americal Society for Information Science and Technology, vol. 61, no. 8, pp: 1517–1534, 2010.

[18] T. Yilmaz et al., "Improving educational web search for question-like queries through subject classification", Information Processing and Management, Vol. 56, Issue. 1, pp. 228–246, 2019.

[19] H. Jing and E. Tzoukermann, "Determining Semantic Equivalence of Terms in Information Retrieval: an Approach Based on Context Distance and Morphology", Recent Advances in Computational Terminology, John Benjamins Publishing Company, January 2001, pp.245-260.

[20] G. Akrivas et al., "Context-Sensitive Semantic Query Expansion", Proceedings of the 2002 IEEE International Conference on Artificial Intelligence Systems (ICAIS'02), p.109, September 2002.

[21] M. Arora et al., "Challenges in Web Information Retrieval", In: Sobh T., Elleithy K. (eds) Innovations in Computing Sciences and Software Engineering. Springer, Dordrecht, pp. 141-146, November 2010.

[22] N. L. M. Shuib, N. Abdullah and M. H. bin Ismail, "The Use of Information Retrieval Tools: a Study of Computer Science Postgraduate Students", 2010 International Conference on Science and Social Research (CSSR 2010), Kuala Lumpur, Malaysia, 2010, pp. 379-384.

[23] M. Sanderson and W. B. Croft, "The History of Information Retrieval Research", Proceedings of the IEEE, vol. 100, no. Special Centennial Issue, pp. 1444-1451, May 2012.

[24] J. Allan et al., "Challenges in Information Retrieval and Language Modeling - Report of a Workshop held at the Center for Intelligent Information Retrieval, University of Massachusetts Amherst, September 2002", ACM SIGIR Forum, vol. 37, no. 1, pp. 31–47, Spring 2003.

[25] J.Allen et al., "Frontiers, Challenges, and Opportunities for Information Retrieval Report from SWIRL 2012 The Second Strategic Workshop on Information Retrieval in Lorne February 2012", ACM SIGI Forum, vol. 46, Issue. 1, pp. 2-32, June 2012.

[26] J. S. Culpepper, F. Diaz, and M. D. Smucker, "Research Frontiers in Information Retrieval: Report from the Third Strategic Workshop on Information Retrieval in Lorne (SWIRL 2018)", ACM SIGIR Forum, Vol. 52 Issue 1, Pages 34-90, June 2018.