TALASH: A SEMANTIC AND CONTEXT BASED OPTIMIZED HINDI SEARCH ENGINE

Nandkishor Vasnik[1], Shriya Sahu[2], Devshri Roy[3]

[1]Department of CSE, MANIT, Bhopal, MP, India
vasnik.nd@gmail.com
[2]Department of CSE, MANIT, Bhopal, MP, India
s.shriya88@gmail.com
[3]Department of CSE, MANIT, Bhopal, MP, India
droy.iit@gmail.com

ABSTRACT

The traditional search engine have shortcoming that they retrieve irrelevant information. Query expansion with relevant words increases the performance of search engines, but finding and using the relevant words is an open problem. This paper presents a Hindi search engine in which we describe three models for query enhancement. They are based on lexical variance, user context and combination of both techniques.

KEYWORDS

Information retrieval, context, lexical resources, query enhancement, HindiWordNet.

1. INTRODUCTION

The Web is growing as the fastest communication medium. This technology in combination with latest electronic storage devices, enable us to keep track of enormous amount of information available to the society. Information present in web are present in different language like English, Arabic, Bengali, Hindi many more. The Web is full of unstructured information and traditional search engines practice a keyword-oriented search scheme which leads to the problem of retrieving numerous irrelevant information. Such searching scheme with a specific keyword may eventuate to unsatisfactory, but with its synonym to appropriate results .Many of the documents retrieved for general queries are irrelevant to the subject of interest and other documents are missing because the query does not include the exact keywords. Users are not confident about the languages used to formulate their queries, refined queries with restrictive Boolean operators may result in a few or even no documents .This motivates the use of natural language interfaces as an adequate way of communicating with search engines.

One solution to improve the relevancy of retrieved results is to expand the user query with more relevant words. But achieve suitable relevant words is a challenging problem. Along with this, sometimes the query is nebulous, its domain or context is unpredictable and as a result, selection of enhancing keywords is really difficult [13]. Instances of these queries are words which have various meanings in different domains.

DOI : 10.5121/ijcseit.2012.2305

To improve the quality of the search on Internet, NLP techniques approaches have typically been adopted. In which query extensions and improving the quality of information retrieved using NLP-based systems. Result of search engine is depend on database present and how structured is it. In web very limited Hindi document is present, so traditional searching scheme will not work properly. We have applied some technique to get better result from limited Hindi document.
Our proposal is to provide linguistic mechanisms that transform and extend the user query by integrating HindiWordNet semantic database, [12] and user context. The main goal of system, TALASH: A Hindi Search Engine is to improve the result provided by Google search engine through the extension of user Hindi input. The rest of the paper is organized as follows. Next section discusses related research in concerned fields. In Section 3, suggest the proposed query expansion models. Section 4 presents Result and experiment and finally, Section 5 gives conclusions and directions for future work.

2. RELATED WORK

According to research accomplished by Imaee and colleagues [1], query expansion has a significant role in the Information retrieval. They showed that it can eventuate to the increase of precision in information retrieval. Since the emergence of Semantic Web in late 1990, semantic search has been one of the most distinguished areas. Strategies for combining information of ontology [2] and electronic dictionary with search patterns are studied [3,4,5]. They show that exploiting only one semantic relation, such as Hypernym or Synonym is not effective, so it is better that a combination of semantic relations to be used. In the studies that a combination of these relations has been used [6, 7], promising results have been reported.

Dey et al. [8] define context 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 their applications. Schilit et al. [9] were one of the first groups to identify context-aware computing as an important aspect of mobile computing. They identify who one is, one's location, and the people, objects, or other resources nearby as important aspects of context. This also aligns well with Bellotti and Edward's classification of necessary types of context. They believe context-aware systems must include responsiveness to the environment, to the person, and to the person's social group [10]. Erickson addresses issues for Context-Aware computing in a CACM article [11]. He discusses how context aware software development often requires the software to act autonomously without a specific request from the user. For example, a software system may automatically adjust the volume of a speaker system in order to compensate for the volume of a speaker's voice. Erickson points out that it is difficult for the software to accurately make decisions without input from the user. Google provide searching in Hindi, Bengali, Tamil and many more but due to unstructured or limited document it retrieve many irrelevant document. Hindi documents present in web are not so huge so use get irrelevant information most of the time.

In Hindi there is limited work has been done on query enhancement to improve the performance of search engine and this motivates for developing enhance Search Engine in which user will pose a query in Hindi and get relevant information related to input.

3. THE PROPOSED QUERY EXPANSION MODEL

We have developed three methods for query enhancement for Hindi input in system. First method use lexical resource, second method use user context [14] and last method is combination of this two method [17].

3.1. Method-I: Query enhancement using lexical resource

Extending web queries using lexical resources focusing on the Query Generator component, it deals with lexical variation of significant words of user queries in order to enhance document searching. Lexical variation is accomplished by using HindiWordNet(HWN) [12], a lexical database that is structured as a top concept ontology that reflects different explicit relationships. Words are organized in synonym sets, called synsets (each synset represents a concept). The synsets are related by hyponym and hypernym (IS-A) relationships.

In Query Generator module the natural language query is tokenized, and keywords that have to be extended by using HWN lexical resources. HWN is used for extracting semantically related terms, in this approach synonym and hypernym words are used[16]. For instance, given the noun "झोपड़ी" (Hut), HindiWordNet provides output shown in figure 1.

Then, nominal and adjectival keywords are then expanded with gender and number variations and verbal keywords with their corresponding infinitive lemma .This step is necessary, as the search engine does not perform any kind of stemming/ morphological inflection for Hindi language. In short, the strategy for Boolean query generation is this: semantically related terms are added to each relevant term connected by ORs [13].

Synset [0] 9718 - NOUN - [झोपड़ी, झोपड़ा, झोंपड़ा, झोंपड़ी, मड़ई, मड़ैया, मड़ाई, मढ़ई, मढ़ा, मढ़ी, आशियाना, आशियाँ]
HYPERNYM : 1901 - NOUN - [घर, गृह, मकान, सदन, शाला, आलय, धाम, निकेतन, निलय, केतन, पण, गेह, सराय, अमा, निषदन, अवसथ, अवस्थान, आगार, आगर, आयतन, आश्रय]
HYPONYM : 18476 - NOUN - [टपरी, टपरा]
ONTO_NODES : भौतिक स्थान (Physical Place)

Figure.1: Output of HindiWordNet

For instance, the term "सूरज" (Sun) is expanded as: (सूरज OR दिवाकर OR भास्कर OR दिनकर OR रवि). Finally, all the sets of expanded terms are connected by the AND operator. For example user query is "सूर्य नमस्कार" (Sun Salutation) is translated into: "(सूर्य ORसूरज OR दिवाकर OR भास्कर OR दिनकर OR रवि) AND (नमस्कार OR नमन OR अभिवादन OR अभिवंदन OR नमस्ते)". This query is ready to be used as input for the web search engine

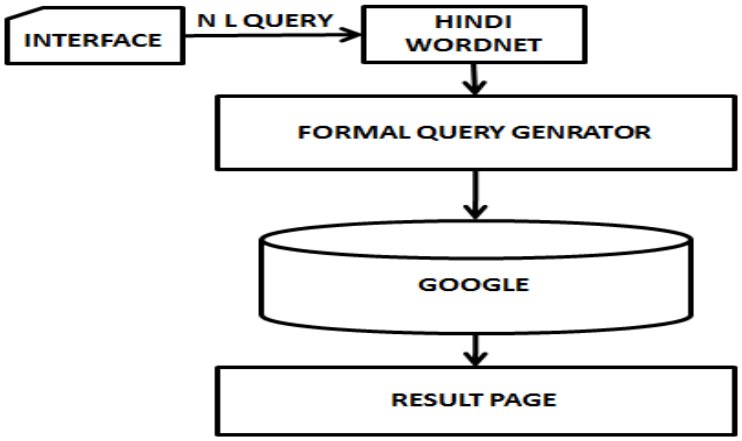

Figure.2: Architecture of Method-I

Architecture for Method-I is shown in figure 2.User impose their Hindi query in system then it pass to HWN. HWN extract semantically related terms and send to Formal Query Generator module. Where Boolean query is builds with the help of semantic related terms and send Google search engine. Search engine retrieve information from web according to Boolean query then show top ten results on interface.

3.2. Method-II: Query Enhancement using user Context

Users are required to register to the system and fill out their profile at the first time they are logging in to the system. In the subsequent times, they are identified by the system. A user can change his profile information in any time he enters the system. The profile information consists of name, username, password, age, profession and sexuality.

3.2.1. Context identification

At first the context information that describe the interaction environment between the user and the system are acquired, which consists of [14]:

(a)User role: User role determines main activity of a user in its current environment. For example a woman is a software engineer then if she searches the query "जावा" (Java) in system, she probably aims "जावा प्रोग्रामिंग भाषा" (programing language) instead of other meanings of "जावा" (java) such as coffee or an island. If same query is fired by a house wife, she expected result on coffee or island not programming language .Hence, user role can be exploited to disambiguate multi-meaning queries.

(b)User location: Location of user will be retrieved from the user's Google calendar. Alternately, the system may collect this information from a local Outlook calendar on the user's local machine. If any user living in particular area then result should be related to that area. If user situated in India and fire query "राष्टीय फूल" (National flower) then result should be related to "कमल" (lotus) national flower of India.

(c)User interests: User interests are entered by users to the system and system augments this by inserting hypernyms of the words [15]. System categorized the user interest in following category and sub categories shown in table 1.Interest of a user is saved for their subsequent search [19]. Users can modify their interest for every new search.

Table 1. Category and subcategory of user interest.

Category	**Subcategory**
शिक्षा (Education)	इतिहास(History),भूगोल(Geography),राजनिति(Politics),मनोविज्ञान (Psychology), संस्कृति(Culture),खगोलिय(Astronomy),दर्शनशास्त्र(Philosophy),समाजशास्त्र(Sociology), साहित्य(Literature).
विज्ञान (Science)	जीव विज्ञान(Zoology), रसायनविज्ञान (Chemistry),भौतिक(Physics),वनस्पति (Botany),प्रोग्रामिंग भाषा(Programming language), कम्प्यूटर हार्डवेयर (Computer hardware),सॉफ्टवेयर(Software), अभियांत्रिकी(Engineering).
मनोरंजन (Entertainment)	फिल्म(Movies),गीत(Songs),छवि(Images), खेल(Sports) ,समाचार(News)
अन्य (Other)	

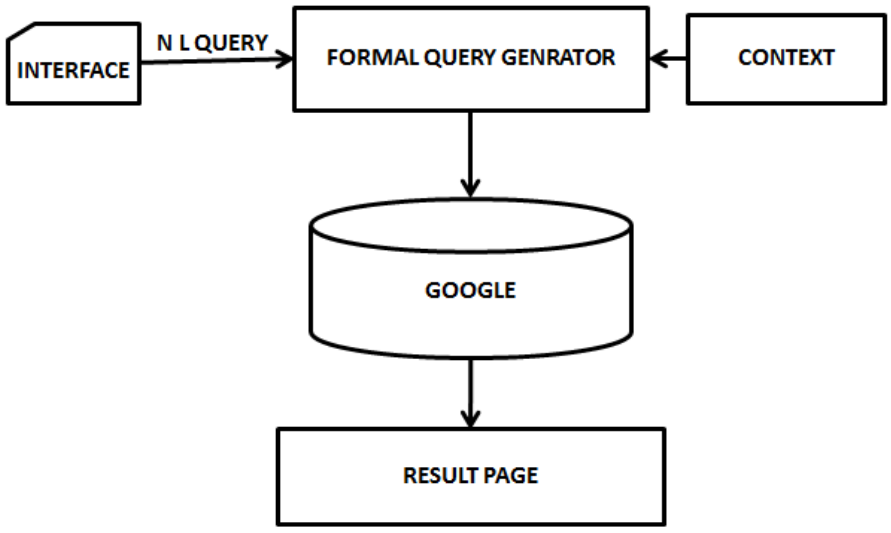

Figure.3: Architecture of Method-II

Architecture for Method-II is shown in figure 3. User login in system through interface and choose their interest. With help of user information and interest formal query generator module form Boolean query and send to Google search engine. Search engine retrieve relevant information from internet according to Boolean query and show top ten results on interface.

3.3. Method-III: Query Enhancement using hybrid technique

Here system uses hybrid technique which is combination of both query enhancement using lexical resource and user Context technique for input hindi query [15]. System builds formal query by submerging query get from both technique.

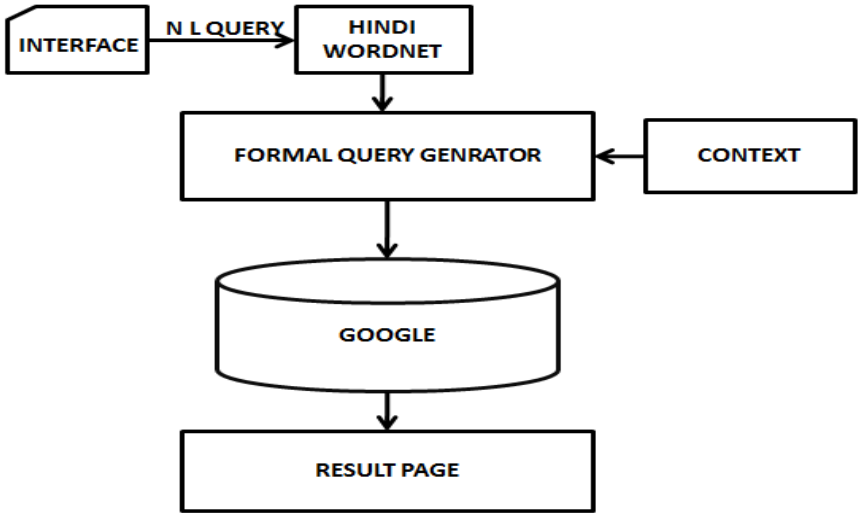

Figure.4: Architecture of Method-III.

Architecture of Method-III is shown in figure 4. User have login in system and impose their query in hindi along with choose one of the category or subcategory of interest. Both semantic related term and user information along with user interest send to formal query generator. It form Boolean Query through received information and send to Google search engine. Search engine retrieve information from web database and top ten results.

4. RESULT AND EXPERIMENT

In order to evaluate how the linguistic knowledge affects the retrieval results, we study of thirty queries in four types of experiments, all of them executed in the Google Search mode. These experiments are: (1) Simple Google search (2) Method-I (query enhance with HWN), (3) Method-II(query enhance with user context) and (4)Method-III (expansion with HWN and user context). In this experiment thirty queries are given one by one to each propose model and analyzed top ten results. Some links are useful and most of results are irrelevant due to insufficient database for Hindi document. Some query give better result as compare to other queries because more Hindi document present in web related to that query. Precision values are measured considering uppermost retrieved documents by experiment [18]. Precision values obtained in the four experiments are: (1) 0.57, (2) 0.67, (3) 0.69 and (4) 0.79. These results show that a combination of Method-I and Method-II could enhance information retrieval.

Table 2.Experiment Results

	Google	Method –I	Method -II	Method -III
Precision	0.57	0.67	0.69	0.79
Relative Recall	0.33	0.49	0.47	0.54
F-Measure	0.418	0.566	0.559	0.641

The recall on the other hand is the ability of a retrieval system to obtain all or most of the relevant documents in the collection. Thus it requires knowledge not just of the relevant and retrieved but also those not retrieved. There is no proper method of calculating absolute recall of search engines as it is impossible to know the total number of relevant in huge databases. So we have adapted the traditional recall measurement for use in the Web environment by giving it a relative flavour. The relative recall value is thus defined as below, where A is total number of document retrieve by search engine and B is sum of document retrieve by all four experiments.

$$\text{Relative Recall} = \frac{A}{B}$$

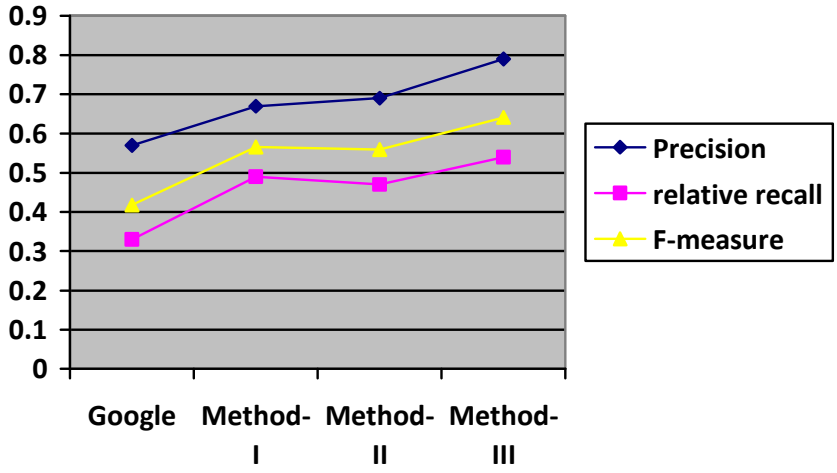

Figure.5: Experimental graph

5. CONCLUSIONS AND FUTURE WORK

In this paper three methods are discuss for query expansion by utilizing HWN, user context and combination of both has been proposed . In first method use semantically related terms of query for this approach synonym and hypernyms words are used. Second method deal with user context, for context, we have investigated the use of personal calendar/PIM information to help determine what a user is currently doing, their physical location and personal interest. Last method implements combination of both methods. The experimental results show that by combining the principles of search with HWN and context awareness, more relevant search results may be produced for a user and, the precision of information retrieval increases. F-

measure also shows that Method-III is better than other methods. Due inadequate Hindi database present in web affect performance of system.

Future direction of the current research consists of designing and implementing agents to extract context automatically by search history of user and implement morphological analyzer like Hindi Shallow Parser with Method-III.

REFERENCE

[1] H. Imai, C. Nigel, and T. Jun'ichi, "A combined query expansion approach for information retrieval" , Genome Informatics, Tokyo, Japan: Universal Academy Press Inc. 1999 [2] J.Davies, R. Studer and P. Warren, Semantic Web Technologies: Trends and Research in Ontology-based Systems, John Wiley & Sons, 2006.

[3] Z. Gong, C. Wa Cheang and U. L. Hou, "Web Query Expansion by WordNet", In Proceedings of the 16th International Conference on Database and Expert Systems Applications ,Copenhagen, Demark, 2005,pp. 166-175.

[4] Z. Gong, C. Wa Cheang and U. L., Hou, "Multi-term Web Query Expansion Using WordNet", the 17th International Conference on Database and Expert Systems Applications,2006.

[5] Y. Liu, C. Li, P. Zhang and Z. Xiong, "A Query Expansion Algorithm based on Phrases Semantic Similarity", International Symposiums on Information Processing, IEEE, 2008.

[6] E. M. Voorhees, "Using WordNet to disambiguate word senses for text retrieval", the 16th Annual international ACM SIGIR Conference on Research and Development in information Retrieval, ACM Press., 1993.

[7] D. Parapar, A. Barreiro, and D. Losada, "Query expansion using Wordnet with a logical model of information retrieval", IADIS International Conference , Portugal, 2005

[8] D. A. Gregory, K. D. Anind, J. B. Peter, D. Nigel, S. Mark, and S. Pete, "Towards a better understanding of context and context-awareness," in Proc. of the 1st international symposium on Handheld and Ubiquitous Computing, Karlsruhe, Germany: Springer-Verlag, 1999.

[9] B. Schilit and M. Theimer, "Disseminating active map information to mobile hosts," IEEE Network, vol. 8, pp. 22-32, 1994.

[10] V. Bellotti and K. Edwards, "Intelligibility and Accountability: Human Considerations in Context-Aware Systems," Journal of Human-Computer Interaction 16, pp. 193-212, 2001.

[11] E. Thomas, "Some problems with the notion of context-aware computing," Communications of the ACM, vol. 45, pp. 102-104, 2002.

[12] http://www.cfilt.iitb.ac.in/wordnet/webhwn/

[13] Ana Garcia-Serrano, Paloma Martine and Albert0 Ruiz "linguistic engineering approach to the enhancement of web-searching" 2001 IEEE.

[14] Najmeh Ahmadian, M .A Nematbakhsh and Hamed Vahdat-Nejad," A Context Aware Approach to Semantic Query Expansion" 2011 International Conference on Innovations in Information Technology,2011 IEEE.

[15] Nicole Anderson," Putting Search in Context: Using Dynamically-Weighted Information Fusion to Improve Search Results", 2011 Eighth International Conference on Information Technology: New Generations,2011 IEEE.

[16] Sírius Thadeu Ferreira da Silva, Samuel de Oliveira Apolonio, Adriana S. Vivacqua, Jonice Oliveira, Geraldo B. Xexéo and Maria Luiza M. Campos," Ontoogle: Enhancing Retrieval with Ontologies and Facets", Proceedings of the 2011 15th International Conference on Computer Supported Cooperative Work in Design,2011 IEEE.

[17] Jingjing Liu, Xiao Li, Alex Acero2 and Ye-Yi Wang," lexicon modeling for query understanding",2011 IEEE.

[18] C. H. C. Leungn and Yuanxi Li," CYC Based Query Expansion Framework for Effective Image Retrieval", 2011 4th International Congress on Image and Signal Processing,2011 IEEE.

[19] Farag Ahmed and Andreas N¨urnberger," A Web Statistics based Conflation Approach to Improve Arabic Text Retrieval", Proceedings of the Federated Conference on Computer Science and Information Systems pp. 3–9,2011 IEEE.

Authors

Nandkishor Vasnik received his B.E. degree in Computer Science & Engineering from Rajeev Gandhi Technical University, Bhopal, India, in 2010. Now he is an MTech student at Computer Science and Engineering Department in Maulana Azad National Institute of Technology, Bhopal, India. His interests involve Natural Language Processing (NLP) and Ontology.

Shriya Sahu received her B.E. degree in Computer Science & Engineering from Chhattisgarh Swami Vivekanand Technical University, Bhilai, India, in 2009. Now she is an MTech student at Computer Science and Engineering Department in Maulana Azad National Institute of Technology, Bhopal, India. Her interests involve Natural Language Processing (NLP) and Ontology.

Dr. Devshri Roy is a University Distinguished Scholar Professor of Computer Science and Engineering at Maulana Azad National Institute of Technology, Bhopal, India . She has done her PhD from Indian Institute of Technology, Kharagpur, India. She is specialized in Application of Computer and Communication Technologies in e-learning , Personalized Information Retrieval,and Natural Language Processing. She published many research papers including writing of books.