International Journal of Electrical and Computer Engineering (IJECE)
Vol. 7, No. 1, February 2017, pp. 432~450
ISSN: 2088-8708, DOI: 10.11591/ijece.v7i1.pp432-450

Context Sensitive Search String Composition Algorithm using User Intention to Handle Ambiguous Keywords

Uma Gajendragadkar[1], Sarang Joshi[2]
[1]COEP, Savitribai Phule Pune University, Pune, Maharshtra, India
[2]PICT, Savitribai Phule Pune University, Pune, Maharshtra, India

Article Info

Article history:

Received Aug 5, 2016
Revised Nov 12, 2016
Accepted Nov 26, 2016

Keyword:

Autocompletion
Context
Data mining
Search
User intention

ABSTRACT

Finding the required URL among the first few result pages of a search engine is still a challenging task. This may require number of reformulations of the search string thus adversely affecting user's search time. Query ambiguity and polysemy are major reasons for not obtaining relevant results in the top few result pages. Efficient query composition and data organization are necessary for getting effective results. Context of the information need and the user intent may improve the autocomplete feature of existing search engines. This research proposes a Funnel Mesh-5 algorithm (FM5) to construct a search string taking into account context of information need and user intention with three main steps 1) Predict user intention with user profiles and the past searches via weighted mesh structure 2) Resolve ambiguity and polysemy of search strings with context and user intention 3) Generate a personalized disambiguated search string by query expansion encompassing user intention and predicted query. Experimental results for the proposed approach and a comparison with direct use of search engine are presented. A comparison of FM5 algorithm with K Nearest Neighbor algorithm for user intention identification is also presented. The proposed system provides better precision for search results for ambiguous search strings with improved identification of the user intention. Results are presented for English language dataset as well as Marathi (an Indian language) dataset of ambiguous search strings.

Corresponding Author:

Uma Gajendragadkar,
COEP,
Phone +919822479128, G7/9 Omkar Garden, Manikbaug, Pune, Maharshtra, India.
Email: umagadkar@gmail.com

1. INTRODUCTION

Current search engines churn a large volume of data to obtain meaningful information; however, the main challenge is to get relevant results in the top few result pages [1], [2]. Search engines check for the presence of keywords in documents. Mere presence of keywords in a document may not match the user's search intention and need. User satisfaction increases when more relevant and exact information is presented in the top few results. An appropriately composed query is the starting point for handling this challenge [3]. Performance of search engines can be improved with the use of appropriate keywords or prediction of such keywords [4-6]. Search engines use search logs and most popular queries; however, these are not sufficient to predict the user's interests or intention [7].

Users are of three types, first - Internet skilled users, second - Internet aware users and third - Internet unskilled users. Many times, users do not know the proper keywords for searching information and they cannot express their information need or intent of search [8], [9]. This results in search results often not satisfying user's information need. This problem can be addressed by query expansion and reformulation [3]. Search engines provide autocompletions of queries based on popularity [10]; however, they are

inadequate [11], [12]. Although different users may use the same query keyword, their intent and context may be different. Current search engines provide the same results to all users using the same keywords at a given point in time. Personalization is desirable to better satisfy the needs of the user [13-15].

The following experiment illustrates this further. If a user searches for 'Michael Jackson' then search engines return results for the famous singer Michael Jackson in majority of result pages. These results would be treated as irrelevant and incorrect if the user intent was to search for professor Michael Jackson.

Table 1. Example search query done on Google [16] on 29th May 2015 result rows

Query String	Michael Jackson	Michael Jackson professor
Total Results	About 39,00,00,000	About 7,89,00,000 results
Search Results as Singer	First 13 pages and after	Page 3 - 5th result
Search Results as Professor	Page 17 8th result	First page
Search Results as Software Development	Page 13 last result	Second page 2nd result
Search Results as VP	Page 16 4th result	Not present in the first 20 pages

As shown in Table 1, when one searches for the query string 'Michael Jackson', results for the singer 'Michael Jackson' are returned in the first 13 pages whereas no result is returned for the professor 'Michael Jackson'. With each page containing 10 results, the relevant results start appearing after 130 result rows. However, when a word 'professor' is added to the query string 'Michael Jackson', the results for professor Michael Jackson are seen in the first result page itself. This demonstrates that if keywords based on user intention are used then better hits can be obtained in the first few search result pages. Query expansion based on user intention has shown to give better search results over large data sets like Web [7], [17].

Thus user intention can be used to disambiguate a query [18]. User context can include parameters such as 'gender', 'age', 'topic', location' etc. It can be short-term [11] or long-term [18].

In the proposed method, user intention is identified with the help of user profile containing parameters like 'gender', 'profession', 'interests', 'location' and past searches. User intention identified with FM5 algorithm is used to reformulate the query. This paper brings together different IR (Information Retrieval) areas like QAC (Query autocompletion), Query Personalization and automatic query expansion.

Our contributions are:

1) A novel user intention identification algorithm is proposed to predict user intention.
2) Query expansion is done using identified user intention to get improved precision for ambiguous search strings.
3) Experimental evaluation of the method is conducted with dataset collected from users. The results reflect improvement in user intention identification and precision of search results.
4) Results of query expansion using the identified user intention are compared with the results of Google search engine [16] directly as first baseline and also with results obtained for ambiguous queries by Chirita et al [7] as a second baseline.

In this paper, Section 2 describes the related work. Section 3 explains data description and how it is used by the proposed system while Section 4 describes the FM5 user intention identification algorithm. Results and discussion are described in Section 5. Conclusion is presented in Section 6.

2. RELATED WORK

2.1. Autocompletions and Personalization

Bhatia et al. [19] presents work where phrases and n-grams are mined from text collections and used for generating autocompletions. Most popular completion i.e. autocompletions based on past popularity of queries in query logs are modeled in Bar-Yossef and Kraus's work [20], [21]. Commercial search engines use MPC (most popular completion) for query autocompletion [20]. Other query autocompletion methods include personalized autocompletion, context based autocompletion using previous queries by user [20], time based autocompletion [22], time and context based autocompletion [21]. Homologous queries and semantically related terms are used to generate autocompletions by Cai et al. [12]

Personalization of query results by using the interests of users is done by many researchers [23-26]. User preferences are collected by either implicit or explicit method. Gender and age are used for personalizing the results by Kharitonov and Serdyukov [27]. User context based on their recent queries is generated and used to rank the query results in a session by Xiang et al [28]. Most of the research conducted is for personalizing the query results by reranking them using user profile rather than query autocompletion.

This paper proposes an algorithm that uses personalization for query completion or autocompletions in search. An improvement in autocompletion ranking is claimed by personalization in Shokouhi's work [29]. Shokouhi et al. also presented ranking of autocompletions with a time-sensitive approach as per their expected popularity [30]. Ambiguous queries are handled by Shoukhoi et al. by providing user context in terms of session context. Query suggestion is achieved by using click information along with previous queries in a session as context and then mining query log sessions for query reformulations [31]. This work is similar to us but it does not consider long-term user context instead focuses on session based user context in terms of click information and previous queries.

2.2. User Intention

Many studies have tried to identify user intention in different ways. Most of them try to categorize the queries as informational, navigational and transactional as proposed by Jansen et al [32]. Given a query suggestion, efforts have been done to understand the user intention using different means like web search logs [26], [33-36], previous user's search log for same query [37], clicked pages [38], user's search session history [39], Wikipedia [40], Wordnet and Google n-gram [41]. Using search query logs for existing users to identify intention cannot guarantee the correctness of search results [37]. Search intent prediction along with query autocompletion is a less explored area. According to Cheng et al., many searches are triggered by browsed web pages [42]. Kong et al. tried to predict search intent using recently browsed news article before search [43]. A large number of queries are triggered by news article daily [43]. Predicting search intent using browsed pages is inadequate [43].Our proposed method uses live RSS newsfeed for query prediction. It makes use of user profiles to predict the search intent.

2.3. Query Expansion

Query expansion is used to reformulate the original user query so as to improve retrieval of search results to better satisfy user needs. One of them is relevance feedback using the returned results and adding new terms related to the original query and selected documents [44]. Other methods include adding relevant terms based on term frequency, document frequency from top ranked documents [45], [46], co-occurrence based techniques [47], thesaurus based techniques [48- 51], desktop specific techniques [7], probability of terms over search logs [52]. Our approach uses a user intention based keyword addition to expand the original query to handle ambiguous query terms.

3. DATA DESCRIPTION

3.1. Data collection Methodology and Data Resources

The system uses different types of data sources. For temporal contextual corpustwo elements are considered. One is static contextual data based on current month and the second is dynamic contextual data based on daily current events. Based on the parameter ‘period', a month-wise list of occasions from Hindu and Christian calendar is taken and their associated keywords list is built. Secondly based on daily current events, RSS news feed from Reuters [54] is processed and a dataset of keywords is built [1]. The temporal data is refreshed every day and also at restart of server. This contextual data is generated for both English and Marathi-an Indian language popularly used in the state of Maharashtra by more than 70 million people. Marathi n-gram dataset is also created by crawling Marathi websites for about four months and processing the web pages and is available [55].

The proposed algorithm also uses data from various sources like Google n-gram [56] and Wordnet [57] for English and Marathi Wordnet data [58]. How to use abovedescribed contextual data to mine possible query autocompletions is discussed by Uma Gajendragadkar et al [1]. Autocompletions for all sample test queries are collected from popular search engines for comparison. This is done foreach character key press of all the test queries.

3.2. User Intention Based Query Expansion

‘K’ user profiles returned by KNN (K Nearest Neighbor) algorithm are used as input to the FM5 algorithm.
Let

$$Z' = \{Z, \bar{Z}\} \quad (1)$$

be a set of profiles such that

$$P(Z') = \sum P(Z) + \sum P(\bar{Z}) \} \quad (2)$$

where $P(Z)$ is the probability of the known user profiles and $P(\bar{Z})$ is the probability of unknown user profiles. Let

$$A = \{a_i | 0 \leq i \leq n-1 \} \quad (3)$$

be the set of *n* query words and let

$$B = \{b_j | 0 \leq j \leq m-1 \} \quad (4)$$

be the set of *m* intentions identified. A trial is conducted by collecting random samples<*ai; bj*>. For each sample query keyword *ai*, there could be multiple user intentions $b_j$stored in intention matrix. If keyword a_i has two possible intentions then they are indicated by 'T' and all other intentions those are not applicable are indicated by 'F'. Total 34 different user intentions are considered. For example consider keyword - 'Jaguar'. It can have two possible user intentions - 'Automobile' and 'Wildlife'. Initially a_i= 0 and b_j= 0 when no query keyword is typed or predicted and hence no user intentions are present.

3.3. Learning Method and Knowledge Generation

Association rule based learning method is used for user intention identification. Support and confidence for an intention are computed for the predicted keyword. Association rule learning is used to find interesting relations between different parameters in the data [59]. It finds strong rules in data based on support and confidence measures. If a rule such as $\{milk, sugar\} \Rightarrow \{coffee\}$is found in the data, it indicates that a customer is likely to buy coffee if the customer has bought both milk and sugar. Association rule is used in many applications like market analysis, bioinformatics, web usage mining etc.

Minimum threshold values on support and confidence are used to find the interesting rules out of all possible rules. If $I = \{i_1, i_2, i_3, \ldots, i_n\}$ is a set of items and $D_b = \{tr_1, tr_2, tr_3, \ldots, tr_m\}$ is a set of transactions in databaseD_b, then a rule can be defined like $U \longrightarrow V$ where $U, V \subseteq$ I and $U \cap V = \phi$.

Support can be calculated as proportion of transactions containing the item set U. Say for illustration, the item set {*milk,sugar*} has a support of 6/10 = 0.6 since it occurs in 60% of all transactions. Confidence of rule $U \longrightarrow V$ can be calculated as the proportion of the transactions that contain both U and V.

$$Conf\{U \longrightarrow V\} = \frac{Supp(U \cap V)}{Supp\ (U)} \quad (5)$$

For illustration, the rule has $\{milk, sugar\} \Rightarrow \{coffee\}$ a confidence of 0.9 means in 90% of the transactions that contain milk and sugar the rule holds true. Other user intention identification methods have few drawbacks. Using web search logs for intent identification lacks in correct outcomes as the same query responses have been provided to the users. Using click pages [38] is not very effective as user clicks do not always translate to the result being relevant to search intent. User search session history [39] works only for a session. No user intention prediction was done for ambiguous query in case of intent identification with Wikipedia [40]. Table 2 shows few records from sample data considered for learning intent of a user for a given keyword.

Table 2. Example Data for Association Rule Mining

Word	Intent	Gender	Location	Profession	Interest
bond	legal	M	India	Lawyer	Cooking
court	legal	M	India	Engineer	Gardening
judge	legal	M	USA	Lawyer	Books
law	legal	F	UK	Farmer	TV
notary	legal	F	India	Doctor	Painting
notice	legal	M	India	Lawyer	Poems
search	legal	F	India	Engineer	Gardening
java	political	F	USA	Doctor	Poems
jaguar	Automobile	M	USA	Engineer	Art
diabetes	health	M	UK	Doctor	Theatre
interest	social	M	India	Farmer	Photography
apple	technological	F	India	Engineer	Wildlife
java	technological	M	India	Engineer	Sports
bond	movie	M	India	Engineer	Movies

From the sample data, all rules having a support and confidence more than the threshold value are considered. These rules are used to learn about the possible intention of a user about a keyword.
Let

$$G = \{g_i \,|1 \leq m\} \tag{6}$$

be the returned intentions and $g_i \epsilon B$ in Equation 4 then how one out of these is selected is explained in section 4 of the paper. For the purpose of experimentation, two types of users are considered for the system - registered user and unregistered user. First case is when a user logs in to the system (registered user with known user profile in Z') and the second case is a user who does not log in to the system (unregistered users with unknown user profiles in$\bar{Z}$) as in Equation 1. In the second case, if user does not login to the system then the user profile is not available hence no personalization can be done and no learning happens. In the first case, a User Profile is created during registration to the search system. This profile 'X_1' is created by obtaining user preferences for a set of questions. The values are filled in by an explicit questionnaire asking questions like 'What is your Profession?' and answer will set the value. User preferences are stored in the user profile and the past searches done by the user are also stored in this profile. A bit vector representing user profile is stored in the system for every registered user. This vector of different parameters forms a key for each user.

X_1 is the set of user profile parameters to be considered for taking decision of next probable alphabet/ numerical/symbol. The composition of search string is done using elements of set X_1 .

$$X_1 = \{X_{10}, X_{11}, X_{12} \ldots., X_{1d}\} \tag{7}$$

The system personalizes search strings based on these user preferences and learns from past searches. After pressing a key character in search box, the system tries to predict the next character by using the past searches of the registered user initially and later by using the pool of searches done by other users having similar profiles to the current user.

Comparison of this method is done with KNN (K Nearest Neighbor) algorithm. The graphs in Figure 1 show the performance of KNN for user intention identification with different *K* values on sample data. As seen in Figure 1, KNN shows better performance with smaller *K* value for identifying the user intention but the accuracy of the identification is less (about 39%). To better predict the user intention, we have proposed the FM5 algorithm.

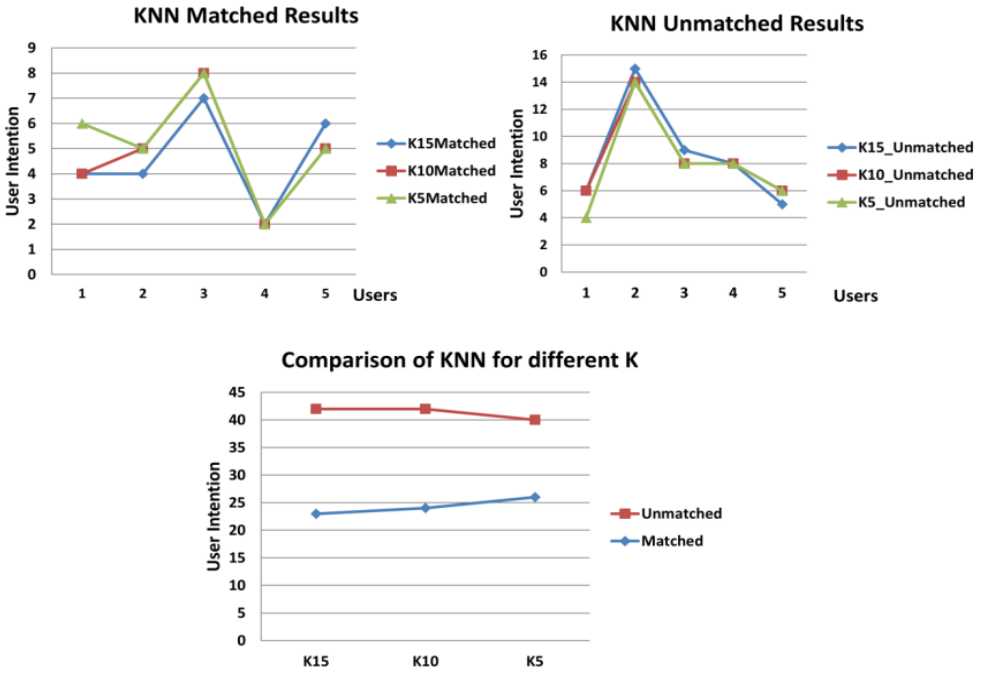

Figure 1. Performance of KNN with Sample Data

4. PROPOSED METHOD

The objective is to find appropriate user intention for a search string being entered by a user in the search box. In FM5 algorithm, user profiles are used to identify user intention for a search string being entered in the search box. For each user a user profile is created during registration to the search system as discussed in section 3. FM5 implements a funnel filter consisting of different meshes mapped to the user profile parameters. Weights are applied to these meshes to disambiguate different user intentions of query word a_i as given in Equation 22. User can select the parameter or multiple parameters to be used. If users' current search intention is related to his {'Interest'} rather than his {'Profession'} then only the parameters like {'Interest, Gender, Location'} could be selected and other parameters like {'Profession'} could be omitted.

Let

$$P_{zi} = \{ P_0 , P_1 , P_2 \ \ldots . , P_d \} \tag{8}$$

be the set of parameters considered for the experiment. For example, user profile consists of 5 parameters- 'Profession', 'Interests', 'Gender', 'Location' and 'Past searches'. For illustration purpose, higher weight is assigned to {'Profession'} parameter followed by {'Gender', 'Interests', 'Location', 'Past searches'} respectively. This is configurable and more parameters can be added to the funnel shown in Figure 2.

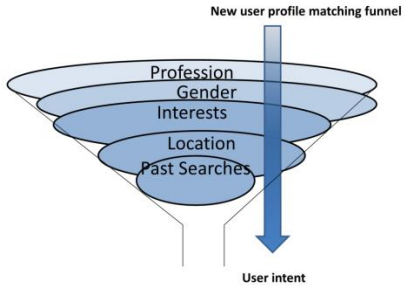

Figure 2. User Intention Identification Funnel

Let

$$W_z = \{ W_0 , W_1 , W_2 \ \ldots . , W_d \} \tag{9}$$

be the set of weights such that

$$f_x : P_{zi} \rightarrow W_z \mid W_i < W_{i+1} \tag{10}$$

The computation of these weights is explained in Equation 22.

Let Q be the prefix query input string which is progressively attached with an alphanumeric character to complete the search string.

$$Q = q_i \mid 0 < n \tag{11}$$

where q_i , q_{i+1}, …. are characters to compose the search string assuming {n} character key presses. Initial state of this set is empty. q_i can be any character ranging from 'a' to 'z' and '0' to '9'or characters like ':, ;, ~, …' etc. Let q_{i+1} is a partial search string.The composition of search string and related selection of q_i are done using elements of set X_1 as per Equation 7.

4.1. Proposed Circular Structure for User Profile Parameters

The user profiles are organized in a circular linked list as shown in Figure 3. A circular linked list is used as one can add or remove parameters from the list easily and easy to traverse to reach an object. Learning will add or subtract parameters from circular linked list. Sizeof the circular linked list will increase or decrease accordingly. Let '*r*' be the radius of the circle on which various user profiles are arranged. The pointer in the circular linked list is placed as per the weight calculated by association rule in terms of support as shown in Equation 20.

Let X_2 be the cost

$$X_2 = C_i \mid 0 \leq C_i \leq 1 \quad (12)$$

associated with elements of Q such that

$$f_1 : Q \rightarrow X_2 \quad (13)$$

C= 1 when partial or complete search string does not exist in the search set or the algorithm fails to predict the search string. C = 0 when search string is distinctly known then no search string prediction and composition is required. The sorting of search string is done such that it always has the central tendency.

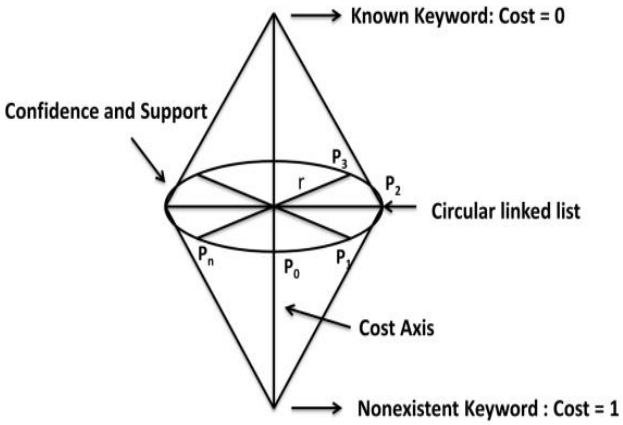

Figure 3. Circular Structure used for FM5 - user Intention Identification Algorithm

Since the search string is computed and mapped in the range [0,1], the central tendency predicts most likely search string. Computation of logical search string is done with the various parameters of user profile.Greedy algorithm is used for intention selection using cost (weight). Each time the mesh with largest cost and greater than or equal to threshold value is selected as explained in Equation 21.

4.2. Mathematical Model for Funnel Mesh 5 (FM5) Algorithm

The algorithm in pseudo code form is represented in the Algorithm 1. Rest of the section explains it in detail.

To illustrate it further, if X_{1i} = *Profession* then

$$X_{1i} = \{\text{Engineer, Doctor, Lawyer, Architect, ..}\} \quad (14)$$

or combination of these. The system assumes that one user has one profession. If 6bits are used to store profession parameter then 2^6 combinations are possible. Forexample, let the bit sequence '000001' indicates *profession = Engineer*.The probability of choosing correct profession is

$$P(X_{1i}) = \frac{1}{2^t} \quad (15)$$

where*t* = number of bits used to store the parameter. Let

$$X_3 = \{ X_{30}, X_{31}, X_{32} \ldots., X_{3n} \} \quad (16)$$

be the past searches associated with the user profile vectors in circular linked list. Let

$$X_{3i} = \{ S_{0,} \ S_{1,} \ S_{2,} \ldots\ldots, S_m \} \quad (17)$$

be the past search strings associated with $X3i$. Then

$$f_2 : S_i \rightarrow W_i \tag{18}$$

where weight Wi is the support value calculated using association rule for search stringSi as shown in Equation 20.
$\forall Si \in X3i$ a list of parameters is given by

$$X_4 = \{ X_{40}, X_{41}, X_{42} \ldots, X_{4k} \} \tag{19}$$

Weight is computed using association rule of the form $X \rightarrow Y$ having value ≥ 0.5as the central tendency is being calculated in circular linked list.

ALGORITHM 1: FM5 User Intention Identification algorithm

U_C = Current user profile;
U_N = Compute nearest User profile vector Matrix;
q_i = Get prefix input;
$K_{w(q_i)}$ = Build set of query strings starting with q_i from past searches $\in U_N$;
for *each* $searchkey \in K_{w(q_i)}$ **do**
  Rules = Compute association rules of type $searchkey \leftarrow parameter/s \in U_N$;
  for *each rule* $R_i \in Rules$**do**
    W = *Compute support* (R_i) *where* $Supp(X \rightarrow Y) = \sigma(X \cup Y)/N$;
    if $W \geq THRESHOLD$ **then**
      end
      P_j = *Get user profile parameter* $\in R_i$;
      X_4 = *Build Mesh* ;
      $Mesh = Mesh \cup P_j$;
    end
    for *each* $P_j \in Mesh$**do**
      for *each user* $\in U_N$ **do**
        if $U_N.P_j = U_C.P_j.value$ **then**
          end
          $X_6 = X_6 \cup user$;
        end
      end
      for *each pastsearh* $\in X_6$ **do**
        if *pastsearch = search key* **then**
          end
          $X_7 = X_7 \cup pastsearch.intention$;
        end
        Return User intention(searchkey) = $\underset{mi \in X_7}{\operatorname{argmax}}$ *f(mi)* ;
        where f(mi) = frequencies of intentions in X_7 ;
      end

Algorithm 1. Funnel Mesh 5 (FM5) Algorithm

Here$\exists X_{1i} \in X$ and $\exists S_i \in Y$ and support value is given by

$$Supp(X \rightarrow Y) = \frac{\sigma (X \cup Y)}{N} \tag{20}$$

where N = total number of records in the circular linked list and X is a combination of parameters from $X1$ whose support ≥ 0.5.
Then, $\forall$currentuser, $\forall S_i$ starting with the prefix Q,

$$X_4 = \{ X_{40}, X_{41}, X_{42} \ldots, X_{4k} \} \mid W_i \geq 0.5 \tag{21}$$

$$W_i = Supp(X \rightarrow Y) \tag{22}$$

$For\ S_i \in X_{3i}$ let

$$X_5 = \{ X_{50}, X_{51}, X_{52} \ldots, X_{5h} \} \tag{23}$$

be the set of possible user intentions.Total 34 intentions are considered for the experimental setup. Table 3 shows user intention taxonomy and example keywords for each of the intentions. For example, the keyword ‘Seed’falls under intention ‘Gardening’ whereas ‘Stanza’ belongs to ‘Poems’.
For a_i as in Equation 3, all search strings starting with prefix Q would be considered. Letthe set of search strings starting with prefix be

$$X_6 = \{ X_{60}, X_{61}, X_{62} \ldots., X_{6k} \} \quad (24)$$

Initial probability of choosing the next character is

$$P(a_i + 1) = \frac{1}{|X_6|} \quad (25)$$

With each character keypressqi+1, this probability increases as $|X_6|$ (count of possible search strings) keeps on reducing as shown in Figure 4. As in Figure 4, with every parameter having weight (support) greater than the threshold, a mesh filtering is done and user intention list keeps on reducing and algorithm makes use of the central tendency to identify matching intentions. From Equation 23, it is seen that there are 'h' intentions. Initially there are 'h' intentions. Hence probability of choosing matching intention will be

$$P(MI) = \frac{1}{h} \quad (26)$$

where MI = matching intention.

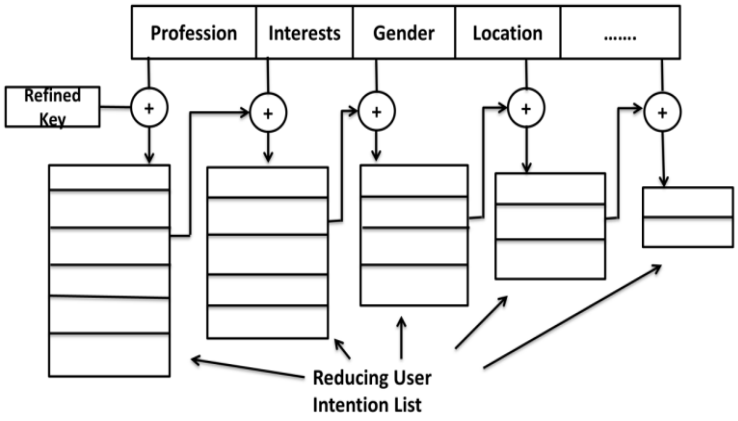

Figure 4. User Intention Identification

For the purpose of experimental trial, total 34 intentions are considered as shown in User intention taxonomy in Table 3. Table 3 shows example keywords for each of the intentions. For example, the keyword ‘seed’ can fall under ‘Agriculture’intention, ‘song' in ‘Music’.

Table 3. User intention taxonomy and examples

Intention	Example	Intention	Example	Intention	Example
Social	diwali	Agriculture	seed	Movies	actress
Technical	java	Bad meaning words	shit	Theater	cast
Research	survey	Music	song	Art	painting
Political	parliament	Cooking	chef	Craft	weaving
Philosophy	thought	Sports	cricket	Painting	color
Medical	liver	Gardening	plants	Travel	destination
Military	arsenal	Health	nutrition	Hiking	harness
Religious	scripture	Books	author	Social work	society
Scientific	experiment	Writing	article	Sculpting	idol
Legal	battle	Poems	poet	Photography	light
New Generation	twitter	TV	show	Literature	novel
				Wildlife	leopard

Based on weights calculated in Equation 22, parameters are chosen from set X_{5i} for each past search string and a further function f_{n3}: $Q \rightarrow MI$ is applied. f_{n3}gives a reduced set of matching intentions from initial {'h'} intentions as shown in Figure 4. Let the returned multiset be given as

$$X_7 = \{X_{70}, X_{71}, X_{72} \ldots., X_{7r}\} \mid r < h \quad (27)$$

So with recursive application of f_{n3} , using elements of X_{5i} , the probability becomes

$$P(MI) = \frac{1}{r} \quad (28)$$

After the last application of f_{n3}, whatever multiset of matching intentions is obtained, from that the final user intention is chosen. For this, multiplicity of each member of the multiset is found and the member with highest multiplicity is chosen as the final user intention.

$$MI = \text{argmax}_{mi \in X7} f(mi) \quad (29)$$

where *f(mi)* = frequencies of intentions in X_7

4.3. Query Expansion

Let *Q* be the original query selected by user.Let *A* be the set of ambiguous queries such that $A \subset Q$.
Let *C* be the set of context based words for each ambiguous word

$$\forall\, a_i \in A, C = \{\, c_j \mid 2 < j < m \,\} \quad (30)$$

where *m* is maximum number of meanings associated with the word a_i.
Query expansion patterns are used to expand the query selected by user based on user intention. Let Q'be the set of query expansion patterns available given by

$$Q' = \{< a_i, c_j > \mid a_i \in A, c_j \in C\} \quad (31)$$

Let

$$f : MI \rightarrow Q' \quad (32)$$

be the function which maps the identified user intention to an expansion pattern from Q'.
The FM5 algorithm identifies the user intention for a query using association rule and user profiles.
The original query is modified as

$$Q = Q \cup Q' \quad (33)$$

and given to search engine.

Table 4. FM5 Intention Identification Examples

Keyword	Cost – Mesh selected	Intentions of final user set after filtering	Matching Intention
Jaguar {User – Female, Engineer, Music, India}	0.6 - Profession → Jaguar 0.7 - Location →Jaguar	Automobile Wildlife Automobile Automobile Automobile	Automobile
Java {User – Female, Engineer, Sports, India}	0.63 - Profession → Java 0.6 - Gender → Java 0.7 - Location → Java	Research Technology Technology	Technology
Bond {User – Male, Engineer, Movies, India}	0.76 - Location → Bond	Movie Legal Movie Movie Legal Movie	Movie

Table 4 lists 3 sample test cases for FM5 algorithm. In the first column, test keyword and the profile of the user entering test keyword is given. In the second column, the association rule (mesh parameter) selected and its cost are given. In the third column, list of possible intentions obtained after filtering through chosen meshes is given. In the fourth column, the matching intention generated as output for the keyword is given. The first keyword Jaguar is entered by a user who is female and an engineer having music as her interest and located in India. After computing the support (cost) for all possible association rules with user profile parameters on LHS and keyword on RHS, only two rules are found having support greater than or equal to threshold of 0.5. Hence two meshes – profession and location are applied. After filtering, a set of 5 users is obtained. From the past searches of these 5 users, intentions for keyword 'Jaguar' are selected and are displayed in the third column. Out of these 5, the most frequently occurring intention – 'Automobile' is returned as matching intention.

5. RESULTS AND DISCUSSION

Authors developed a questionnaire to collect the user profiles and the desired intents for the search strings as shown in Figure 5. For English dataset-1, 25 users and 15 ambiguous queries per user and their desired intent for each query were collected. Thus, 375 queries and intentions were evaluated for the first dataset. For English dataset-2, 100 users and 20 ambiguous queries per user and their desired intent for each query were collected. For English dataset-2 overall 2000 queries and intentions were evaluated. For Marathi dataset, 20 users and 18 ambiguous queries per user were evaluated. Thus, Marathi dataset contained overall 360 queries and intentions. The survey was designed as a paper-and-pencil-based field survey to approach a large number of users and a digital survey was also designed on the same lines. The paper-based questionnaire was designed in two languages i.e. English and Marathi.To validate the proposed model, the questionnaire was distributed to collect the user profile information and desired intention while searching for various ambiguous queries. Population of the study comprises of Engineers, Doctors, Farmers and Lawyers. Samples of 170 users were selected randomly. After scrutiny of filled questionnaire 150 were found to be fit for the analysis. The users comprised of third year Engineering students from different streams of College of Engineering Pune [60] as well as engineers, doctors, farmers and lawyers from different locations in India.

USER INTENTION SURVEY

1. **What is your name?**

   _____ _____ _____

2. **What is your gender?**

   O Male O Female

3. **What is your profession?**

   O Engineer O Doctor O Farmer O Lawyer

4. **What is your best matching interest among the following?**

   O Music O Cooking O Sports O Gardening O Health O Books
   O Writing O Poems O Movie O Theatre O Art O Craft
   O Painting O Travel O Hiking O Social work O Photography O Literature
   O Wildlife

5. **What is your country of residence?**

   _____

6. **Please select your intention for each keyword**

Intention→ Keyword ⇩	Technology	Agriculture	Medical	Political	Research	Automobile	Wildlife	Movie	Military	Legal	Cooking	Health	Music	Finance	Scientific	Social
Apple	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Jaguar	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Java	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Bond	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Windows	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Appendix	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Exercise	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Interest	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Marriage	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Bank	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Seal	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Paint	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Michael Jackson	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Eclipse	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Sun	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O

7. **Specialization:** _____

Figure 5. User Survey Questionnaire designed to collect user profile and intentions

The system is evaluated for 25 users for English dataset-1 with about 15 ambiguous queries each. The training data consists of about 55 user profiles and their past searches. Table 6 shows user intention identification results for FM5 algorithm and KNN algorithm with different '*k*' values like 5 (KNN5), 10 (KNN10) and 15 (KNN15). Matched intention indicates the total number of ambiguous test queries for all test users where the algorithm gave matching intention to the desired intention of user. Unmatched intentions indicate the number of cases where algorithm failed to identify desired intention. The results for user intention identification, obtained with FM5 are encouraging. For English ambiguous dataset-1, accuracy of about 75% is observed with FM5 whereas with KNN an accuracy of about 38.4% is observed for KNN5 and 29.8% with KNN10 and 27% with KNN15.

Table 6. English Ambiguous Dataset Results

Intention/Method	FM5	KNN15	KNN10	KNN5
Matched	282	102	112	144
Unmatched	93	273	263	231
Total	375	375	375	375
Accuracy	0.752	0.272	0.298667	0.384

The first graph in Figure 6 shows Matched intentions obtained for FM5 and KNN for different users with various queries on English dataset. 'Matched' is the legend used for cases where appropriate user intention is obtained and 'Unmatched' is the legend showing cases where the algorithm failed to identify appropriate user intention. The second graph in Figure 6 shows comparison of FM5 with KNN algorithm for different values of '*k*'. FM5 gives better results than KNN.

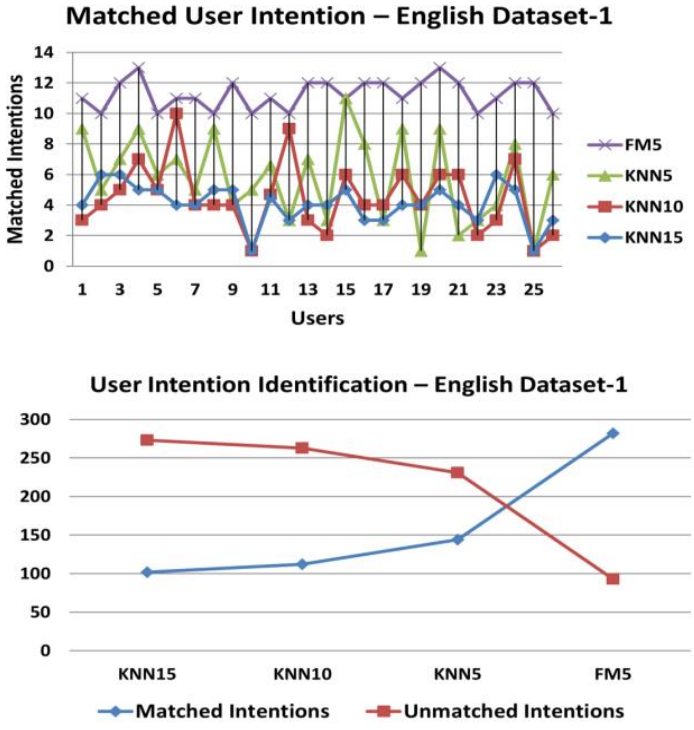

Figure 6. Results for Ambiguous English Dataset-1

For Marathi ambiguous dataset evaluation, system is evaluated with 20 users with 18 ambiguous queries each as shown in Table 7. Thus, 360 queries and intentions are evaluated for Marathi dataset. Accuracy of about

77.5% is observed with FM5 algorithm whereas with KNN an accuracy of about 36.7% is observed for KNN5 and 28.3% with KNN10 and 24.2% with KNN15.

Table 7. Marathi Ambiguous Dataset Results

Intention/Method	FM5	KNN15	KNN10	KNN5
Matched	279	87	102	132
Unmatched	81	273	258	228
Total	360	360	360	360
Accuracy	0.775	0.242	0.283	0.366

User intention identification for search string with FM5 algorithm gives encouraging results. Figure 7 depicts the results for Marathi dataset. The first graph in Figure 7 shows the total number of Matched intentions obtained with FM5 and KNN for each user. The second graph in Figure 7 shows the total number of matched and unmatched intentions for all users with FM5 and KNN.

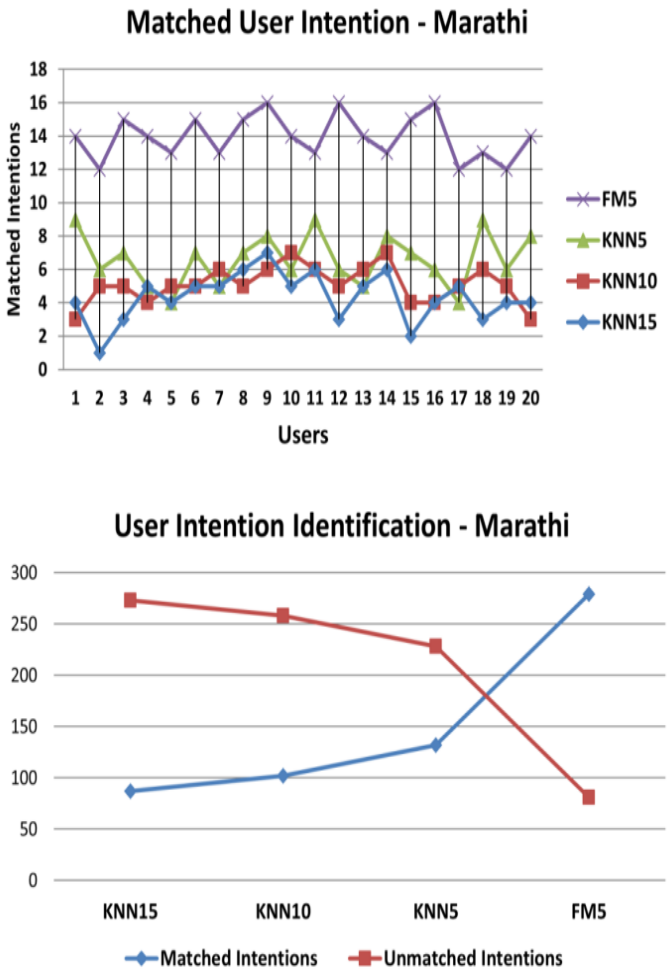

Figure 7. Results for Ambiguous Marathi Dataset

The accuracy observed with FM5 and KNN algorithm for English dataset is plotted in Figure 8.
The accuracy observed with FM5 and KNN algorithm for Marathi dataset is plotted in Figure 9.

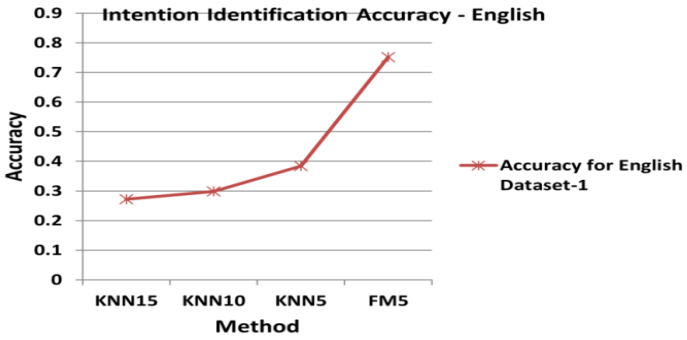

Figure 8. Accuracy observed with FM5 User Intention Identification for English

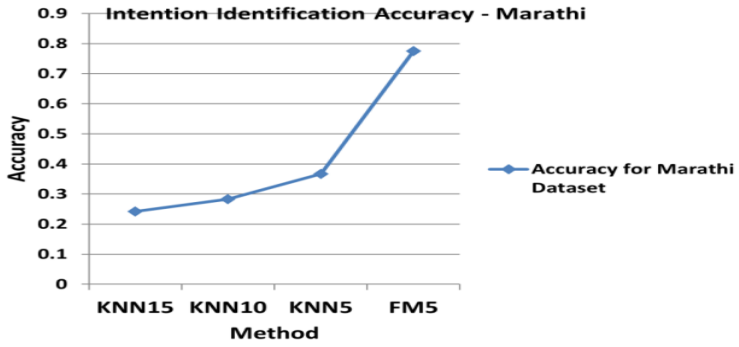

Figure 9. Accuracy observed with FM5 User Intention Identification for Marathi

Further testing of the FM5 algorithm was done using English ambiguous dataset-2. The system is further evaluated for 100 users for English dataset-2 with 20 ambiguous queries each as discussed earlier. Table 8 shows user intention identification results for FM5 algorithm and KNN algorithm with different '*k*' values like 5 (KNN5), 10 (KNN10) and 15 (KNN15). Matched intention indicates the total number of ambiguous test queries for all test users where the algorithm gave matching (correct) intention to the desired intention of user. Unmatched intentions indicate the number of cases where algorithm failed to identify desired intention. The results obtained for English dataset-2 with FM5 show improvement of 0.7% as compared to English dataset-1. This may be because of increased number of past searches available while computing the parameters to build the mesh. For English ambiguous dataset-2, accuracy of a 75.9% is observed with FM5 whereas with KNN an accuracy of about 39.3% is observed for KNN5 and 29.05% with KNN10 and 28.3% with KNN15.

Table 8. English Ambiguous Dataset-2 Results

Intention/Method	FM5	KNN15	KNN10	KNN5
Matched	1518	566	581	786
Unmatched	482	1434	1419	1214
Total	2000	2000	2000	2000
Accuracy	0.759	0.283	0.2905	0.393

The first graph in Figure 10 shows Matched intentions obtained for FM5 and KNN for different users with various queries on English dataset-2. 'Matched' is the legend used for cases where appropriate user intention is obtained and 'Unmatched' is the legend showing cases where the algorithm failed to identify

appropriate user intention. The second graph in Figure 10 shows comparison of FM5 with KNN algorithm for different values of '*k*' for English dataset-2.

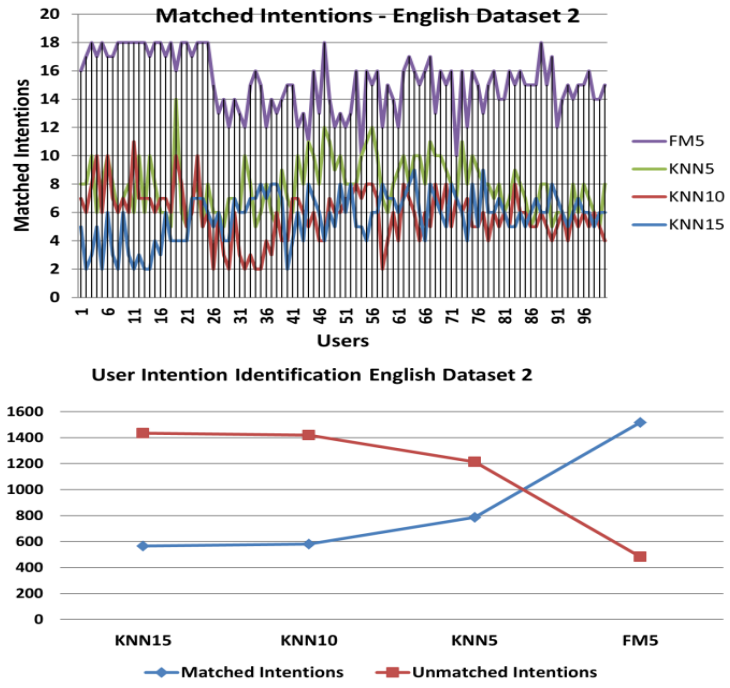

Figure 10. Results for Ambiguous English Dataset-2

Table 9 shows the results of query expansion after user intention identification is done. Top 50 URLs returned by Google API [16] were collected for each query after query expansion. The web pages of these URLs were evaluated as either relevant or not relevant to the query under consideration. Table 9 shows average precision values obtained for test queries after query expansion based on identified user intention.

Table 9. Average Precision after Query Expansion

Query Results	Average Precision
Top 5	1
Top 10	0.986
Top 15	0.982
Top 20	0.984
Top 25	0.983
Top 30	0.98
Top 35	0.97
Top 40	0.964
Top 45	0.96
Top 50	0.955

Metric used for evaluation of search results returned after query expansion is

$$P@N = related\ queries / N \quad (34)$$

It indicates how many valuable results are present in top *N* search results.

5.1. Precision Improvement with FM5 and Query Expansion

Top 50 URLs returned by Google were collected for each query. The results were evaluated as either relevant or not relevant. Precision was calculated. It is given as

$$P = (Relevant\ results) / (Returned\ Results) \quad (35)$$

Graph in Figure 11 shows precision values for top 5, top 10 up to top 50 results obtained with our system and with results from direct use of Google search engine. On X-axis, 1, 2, 3,.. indicate the average precision values for all the queries given by the user for the top 5 search results, for top 10 search results, for top 15 search results, … etc.

From the observed values, the system shows significant improvement in the average precision values. The results are compared to the results obtained using Google search engine [16] and the results obtained for ambiguous queries by Chirita et.al [7]. About 60.47% improvement is seen in average precision with FM5 and query expansion as compared to the results obtained using Google [16]. This is our first baseline comparison. Second baseline comparison is with the results obtained by Chirita et.al [7] and an improvement of 40% is observed with the proposed method.

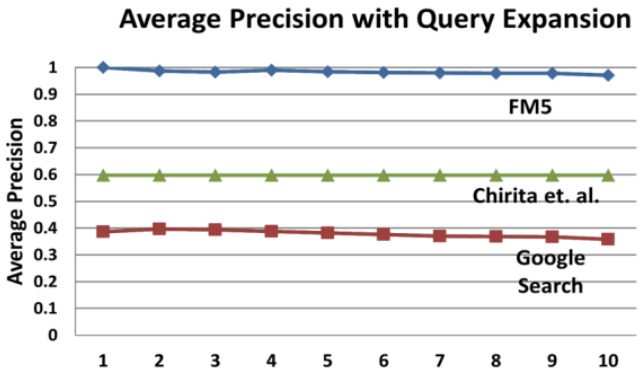

Figure 11. Improvement in Precision after Query Expansion

6. CONCLUSION

Composing the search string by providing better autocompletions to the user that will result in more relevant and less redundant results is the goal of this research. In this algorithm, personalization is used to add context and user intention to the search string composition. The algorithm selects the most appropriate intention out of possible intentions for a keyword by using support as weight. The FM5 algorithm of intent identification via funneling builds upon the advantages of simple k-nearest neighbor (KNN) algorithm.

The approach consists of identification of user intention with FM5 and then expanding original the query based on this intention to obtain more relevant search results in the first few pages. FM5 user intention identification algorithm uses association rule mining with user profiles and shows improvement in performance as compared to KNN. This FM5 when extended with query expansion patterns shows improvement in average precision values for ambiguous queries giving better search results. The system does not use explicit feedback or other strategies like using click pages or session history for determining user intention or for query expansion. Proposed user intention identification algorithm - FM5 showed improvement in accuracy as compared to KNN.

Proposed query expansion approach using identified user intention with FM5 showed improvement in average precision values for ambiguous queries giving better search results in top 50 pages. Experimental results for the proposed approach and a comparison with direct use of search engine showed that performance improved significantly. The proposed system provides better precision for search results for ambiguous search strings with improved identification of the user intention for English language dataset as well as Marathi (an Indian language) dataset of ambiguous search strings.

REFERENCES

[1] U. Gajendragadkar and S. Joshi, "User intended context sensitive mining algorithm for search string composition", in *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent AgentTechnology*, WI-IAT 2015, Singapore, December 6-9, 2015 - Volume I, 2015, pp. 233–236. [Online]. Available: http://dx.doi.org/10.1109/WI-IAT.2015.212

[2] A.A.S. Mamoon H. Mamoon, Hazem M. El-Bakry, "Visualization for information retrieval based on fast search technology", *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, vol. 1, no. 1, pp. 27–42,Mar. 2013.

[3] L. Bing, W. Lam, T.L. Wong, and S. Jameel, "Web query reformulation via joint modeling of latent topicdependency and term context", *ACM Trans. Inf. Syst.*, vol. 33, no. 2, pp. 6:1–6:38, Feb. 2015. [Online]. Available: http://doi.acm.org/10.1145/2699666
[4] G. Salton and C. Buckley, "Improving retrieval performance by relevance feedback", in *Journal of the American Society for Information Science*, 1990, pp. 288–297.
[5] V. Lavrenko and W. B. Croft, "*Relevance-based language models*", in Proceedings of SIGIR '01. New York, NY, USA: ACM, 2001.
[6] M. Harvey, C. Hauff, and D. Elsweiler, "*Learning by example: Training users with high-quality querysuggestions*", in Proceedings of the SIGIR '15, ser. SIGIR '15. New York, NY, USA: ACM, 2015, pp.133–142. [Online]. Available: http://doi.acm.org/10.1145/2766462.2767731
[7] P.A. Chirita, C.S. Firan, and W. Nejdl, "*Personalized query expansion for the web*", in Proceedingsof the 30th Annual International ACM SIGIR Conference on Research and Development in InformationRetrieval, ser. SIGIR '07. New York, NY, USA: ACM, 2007, pp. 7–14. [Online]. Available: http://doi.acm.org/10.1145/1277741.1277746
[8] A. Spink, H. Greisdorf, and J. Bateman, "From highly relevant to not relevant: Examining differentregions of relevance", *Inf. Process. Manage*, vol. 34, no. 5, pp. 599–621, Sep. 1998. [Online]. Available: http://dx.doi.org/10.1016/S0306-4573(98)00025-9
[9] S.D. Torres and I. Weber, "*What and how children search on the web*", in Proceeding CIKM '11. New York, USA: ACM, 2011, pp. 393–402. IJECE Vol. x, No. x, October 2016: 1 – 20IJECE ISSN: 2088-8708 17
[10] S. Chelaru, I.S. Altingovde, S. Siersdorfer, and W. Nejdl, "Analyzing, detecting, and exploitingsentiment in web queries", *ACM Trans. Web*, vol. 8, no. 1, pp. 6:1–6:28, Dec. 2013. [Online]. Available: http://doi.acm.org/10.1145/2535525
[11] H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen, and H. Li, "*Context-aware query suggestion by mining clickthroughand session data*", in KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, NY, USA: ACM, 2008, pp. 875–883.
[12] F. Cai and M. de Rijke, "Learning from homologous queries and semantically related terms for query autocompletion", *Information Processing & Management*, 2016, (Online January 12, 2016.).
[13] D. Sullivan, "The older you are, the more you want personalized search", 2004. [Online]. Available: http://searchenginewatch.com/searchday/article.php/3385131
[14] M.R. Ghorab, D. Zhou, A. O'connor, and V. Wade, "Personalised information retrieval: Survey andclassification", *User Modeling and User-Adapted Interaction*, vol. 23, no. 4, pp. 381–443, Sep. 2013. [Online]. Available: http://dx.doi.org/10.1007/s11257-012-9124-1
[15] A.I.S. Shereen H Ali, Ali I El Desouky, "A new profile learning model for recommendation system basedon machine learning technique", *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, vol. 4, no. 1, pp. 81–92, Mar. 2016.
[16] Google, "Google," 2016, http://www.google.com.
[17] Z. Liu, S. Natarajan, and Y. Chen, "*Query expansion based on clustered results*", Proc. VLDB Endow, vol. 4, no. 6, pp. 350–361, Mar. 2011. [Online]. Available: http://dx.doi.org/10.14778/1978665.1978667
[18] F. Liu, C. Yu, and W. Meng, "*Personalized web search by mapping user queries to categories*", in Proceedingsof the Eleventh International Conference on Information and Knowledge Management, ser. CIKM '02. NewYork, NY, USA: ACM, 2002, pp. 558–565. [Online]. Available: http://doi.acm.org/10.1145/584792.584884
[19] S. Bhatia, D. Majumdar, and P. Mitra, "*Query suggestions in the absence of query logs*", in Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, ser. SIGIR '11. New York, NY, USA: ACM, 2011, pp. 795–804. [Online]. Available:http://doi.acm.org/10.1145/2009916.2010023
[20] N.K. Ziv Bar-Yossef, "*Context-sensitive query auto-completion*", in Proceedings of the 20th International Conferenceon World Wide Web, WWW 2011, Hyderabad, India, 2011.
[21] F. Cai, S. Liang, and M. de Rijke, "*Time-sensitive personalized query auto-completion*", in Proceedings ofthe CIKM '14, ser. CIKM '14. New York, NY, USA: ACM, 2014, pp. 1599–1608. [Online]. Available: http://doi.acm.org/10.1145/2661829.2661921
[22] S. Whiting and J. M. Jose, "*Recent and robust query auto-completion*", in Proceedings of the 23rd International Conference on World Wide Web, ser. WWW '14. New York, NY, USA: ACM, 2014, pp. 971–982. [Online]. Available: http://doi.acm.org/10.1145/2566486.2568009
[23] P A. Chirita, W. Nejdl, R. Paiu, and C. Kohlsch¨utter, "*Using odp metadata to personalize search*", in Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ser. SIGIR '05. New York, NY, USA: ACM, 2005, pp. 178–185. [Online]. Available: http://doi.acm.org/10.1145/1076034.1076067
[24] X. Shen, B. Tan, and C. Zhai, "*Implicit user modeling for personalized search*", in Proceedings of the 14th ACM International Conference on Information and Knowledge Management, ser. CIKM '05. New York, NY, USA: ACM, 2005, pp. 824–831. [Online]. Available: http://doi.acm.org/10.1145/1099554.1099747
[25] Z. Dou, R. Song, and J.-R. Wen, "*A large-scale evaluation and analysis of personalized search strategies*", in Proceedings of the 16th International Conference on World Wide Web, ser. WWW '07. New York, NY, USA: ACM, 2007, pp. 581–590. [Online]. Available: http://doi.acm.org/10.1145/1242572.1242651
[26] J. Teevan, S.T. Dumais, and D.J. Liebling, "*To personalize or not to personalize: Modeling queries withvariation in user intent*", in Proceedings of SIGIR 08, ser. SIGIR '08. New York, NY, USA: ACM, 2008, pp. 163–170. [Online]. Available: http://doi.acm.org/10.1145/1390334.1390364

[27] E. Kharitonov and P. Serdyukov, "*Demographic context in web search re-ranking*", in Proceedings of the 21st ACM International Conference on Information and Knowledge Management, ser. CIKM '12. New York, NY, USA: ACM, 2012, pp. 2555–2558. [Online]. Available: http://doi.acm.org/10.1145/2396761.2398690
[28] B. Xiang, D. Jiang, J. Pei, X. Sun, E. Chen, and H. Li, "*Context-aware ranking in web search*", in Proceedings of the 33rd International ACM SIGIR Conference on Research and Development inInformation Retrieval, ser. SIGIR '10. New York, NY, USA: ACM, 2010, pp. 451–458. [Online]. Available:http://doi.acm.org/10.1145/1835449.1835525
[29] M. Shokouhi, "*Learning to personalize query auto-completion*", in Proceeding of SIGIR '13. New York, NY, USA: ACM, 2013, pp. 103–112.
[30] M. Shokouhi and K. Radinsky, "*Time-sensitive query auto-completion*", in Proceedings of SIGIR'12, ser. SIGIR '12. New York, NY, USA: ACM, 2012, pp. 601–610. [Online]. Available:http://doi.acm.org/10.1145/2348283.2348364
[31] M. Shokouhi, M. Sloan, P.N. Bennett, K. Collins-Thompson, and S. Sarkizova, "*Query suggestion and data fusion in contextual disambiguation*", in Proceedings of the 24th International Conference on World Wide Web, ser.WWW'15. Republic and Canton of Geneva, Switzerland: InternationalWorldWideWeb Conferences SteeringCommittee, 2015, pp. 971–980. [Online]. Available: http://dl.acm.org/citation.cfm?id=2736277.2741646
[32] B.J. Jansen and D. Booth, "*Classifying web queries by topic and user intent*", in CHI '10 Extended Abstracts onHuman Factors in Computing Systems, ser. CHI EA '10. New York, NY, USA: ACM, 2010, pp. 4285–4290. [Online]. Available: http://doi.acm.org/10.1145/1753846.1754140
[33] R. Baeza-Yates, L. Calder´on-Benavides, and C. Gonz´alez-Caro, "*The intention behind web queries*", in Proceedings of the 13th International Conference on String Processing and Information Retrieval, ser. SPIRE'06. Berlin, Heidelberg: Springer-Verlag, 2006, pp. 98–109.
[34] M. Strohmaier and M. Kr¨oll, "Acquiring knowledge about human goals from search query logs", *Inf. Process. Manage.*, vol. 48, no. 1, pp. 63–82, Jan. 2012. [Online]. Available: http://dx.doi.org/10.1016/j.ipm.2011.03.010
[35] D. Jiang, J. Pei, and H. Li, "Mining search and browse logs for web search: A survey", *ACM Trans. Intell. Syst. Technol.*, vol. 4, no. 4, pp. 57:1–57:37, Oct. 2013. [Online]. Available: http://doi.acm.org/10.1145/2508037.2508038
[36] A. Kathuria, B. J. Jansen, C. Hafernik, and A. Spink, "Classifying the user intent of web queries using k-means clustering", in *Internet Research*, Vol. 20 Iss: 5, pp.563 - 581, 2010.
[37] K. Park, H. Jee, T. Lee, S. Jung, and H. Lim, "Automatic extraction of user's search intention fromweb search logs", *Multimedia Tools Appl.*, vol. 61, no. 1, pp. 145–162, 2012. [Online]. Available:http://dx.doi.org/10.1007/s11042-010-0723-8
[38] A. Das, C. Mandal, and C. Reade, "*Determining the user intent behind web search queries by learning frompast user interactions with search results*", in Proceedings of the 19th International Conference on Managementof Data, ser. COMAD '13. Mumbai, India, India: Computer Society of India, 2013, pp. 135–138. [Online]. Available: http://dl.acm.org/citation.cfm?id=2694476.2694508
[39] J.M. Fern´andez-Luna, J.F. Huete, and J.C. Rodr´ıguez-Cano, "*User intent transition for explicit collaborative search through groups recommendation*", in Proceedings of the 3rd International Workshop on Collaborative Information Retrieval, ser. CIR '11. New York, NY, USA: ACM, 2011, pp. 23–28. [Online]. Available: http://doi.acm.org/10.1145/2064075.2064083
[40] J. Hu, G. Wang, F. Lochovsky, J.-t. Sun, and Z. Chen, "*Understanding user's query intent with wikipedia*", in Proceedings of the 18th International Conference on World Wide Web, ser. WWW '09. New York, NY, USA: ACM, 2009, pp. 471–480. [Online]. Available: http://doi.acm.org/10.1145/1526709.1526773
[41] M. Hwang, D.H. Jeong, J. Kim, S.-K. Song, and H. Jung, "Activity inference for constructing user intention model", *Comput. Sci. Inf. Syst.*, vol. 10, pp. 767–778, 2013.
[42] Z. Cheng, B. Gao, and T.Y. Liu, "*Actively predicting diverse search intent from user browsing behaviors*", in Proceedings of the 19th International Conference on World Wide Web, ser. WWW '10. New York, NY, USA: ACM, 2010, pp. 221–230. [Online]. Available: http://doi.acm.org/10.1145/1772690.1772714
[43] W. Kong, R. Li, J. Luo, A. Zhang, Y. Chang, and J. Allan, "*Predicting search intent based on pre-search context*", in Proceedings of the 38th International ACM SIGIR Conference on Research and Development inInformation Retrieval, ser. SIGIR '15. New York, NY, USA: ACM, 2015, pp. 503–512. [Online]. Available: http://doi.acm.org/10.1145/2766462.2767757
[44] J.J. Rocchio, "*Relevance feedback in information retrieval*", in The Smart retrieval system - experiments inautomatic document processing, G. Salton, Ed. Englewood Cliffs, NJ: Prentice-Hall, 1971, pp. 313–323.
[45] A.M. Lam-Adesina and G.J.F. Jones, "*Applying summarization techniques for term selection in relevance feedback*", in Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ser. SIGIR '01. New York, NY, USA: ACM, 2001, pp. 1–9. [Online]. Available: http://doi.acm.org/10.1145/383952.383953
[46] J. Xu and W.B. Croft, "*Query expansion using local and global document analysis*", in Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ser. SIGIR '96. New York, NY, USA: ACM, 1996, pp. 4–11. [Online]. Available:http://doi.acm.org/10.1145/243199.243202
[47] M.C. Kim and K.S. Choi, "A comparison of collocation-based similarity measures in query expansion", *Inf. Process. Manage*, vol. 35, no. 1, pp. 19–30, Jan. 1999. [Online]. Available: http://dx.doi.org/10.1016/S0306-4573(98)00040-5
[48] G. Miller, "Wordnet: An electronic lexical database", in *Communications of the ACM*, vol. 38(11), 1995, p. 3941.

[49] C. Shah and W.B. Croft, "*Evaluating high accuracy retrieval techniques*", in Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ser. SIGIR '04. New York, NY, USA: ACM, 2004, pp. 2–9. [Online]. Available: http://doi.acm.org/10.1145/1008992.1008996
[50] S. Kim, H. Seo, and H. Rim, "*Information retrieval using word senses: root sense tagging approach*", in SIGIR 2004: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK, July 25-29, 2004, 2004, pp. 258–265. [Online]. Available: http://doi.acm.org/10.1145/1008992.1009038
[51] D. Fogaras and B. R´acz, "*Scaling link-based similarity search*", in Proceedings of the 14th International Conference on World Wide Web, ser. WWW '05. New York, NY, USA: ACM, 2005, pp. 641–650. [Online]. Available: http://doi.acm.org/10.1145/1060745.1060839
[52] H. Cui, J.R. Wen, J.Y. Nie, and W.Y. Ma, "*Probabilistic query expansion using query logs*", in Proceedings of the 11th International Conference on World Wide Web, ser. WWW '02. New York, NY, USA: ACM, 2002, pp.325–332. [Online]. Available: http://doi.acm.org/10.1145/511446.511489
[53] E. Di Buccio, M. Melucci, and F. Moro, "Detecting verbose queries and improving information retrieval", *Inf. Process. Manage.*, vol. 50, no. 2, pp. 342–360, Mar. 2014. [Online]. Available: http://dx.doi.org/10.1016/j.ipm.2013.09.003
[54] Reuters, "*Reuters news feed*", 2014, http://feeds.reuters.com/reuters/.
[55] U. Gajendragadkar and S. Joshi, "Marathi language n-gram dataset", in *Indian Language Technology Proliferation and Deployment Centre*, 2015. [Online]. Available: "http://www.tdildc.in/index.php?option=com downloadtask¯showresourceDetailstoolid¯1644lang¯en"
[56] P. Norvig, *Natural Language Corpus Data from Beautiful Data*. Segaran and Hammerbacher, 2009.
[57] wordnet.princeton.edu, "Wordnet: http://wordnet.princeton.edu/," March 2006, http://wordnet.princeton.edu/.
[58] M. IITB, "*Marathi wordnet*", 2014, http://www.cfilt.iitb.ac.in/wordnet/webmwn/.
[59] R. Agrawal, T. Imieli´nski, and A. Swami, "*Mining association rules between sets of items in large databases*", in Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, ser. SIGMOD '93. New York, NY, USA: ACM, 1993, pp. 207–216. [Online]. Available: http://doi.acm.org/10.1145/170035.170072
[60] C. of Engineering Pune, "Coep," 2016, http://www.coep.org.in.

BIOGRAPHIES OF AUTHORS

Uma Gajendragadkar is a PhD Scholar at COEP, Savitribai Phule Pune University, Pune, Maharashtra, India. She completed Masters in Computer Engineering from Mumbai University, India in 2004 and Bachelors in Electronics Engineering in 1993 from Shivaji University, Kolhapur, India. She worked as a TEQIP Research Fellow for past 4 years at COEP, SPPU, Pune, Maharashtra, India. She has 8 years of experience in Software Industry and 13 years of experience in Academics where she taught Undergraduate and Postgraduate Engineeing students. She is member of IEEE and ACM for past eight years.

Dr. Sarang Joshiis a Professor at PICT, SavtribaiPhule Pune University, Pune, Maharashtra, India. He completed his PhD in Computer Science and Engineering from Bharati Vidyapeeth, Pune, India. He completed Masters in Computer Engineering and Bachelors in Computer Engineering from University of Pune, India. He works as a Professor in Computer Engineering at PICT, SPPU, Pune, Maharshtra, India for last 27 years. He was the Chairman of Board of Studies of Computer Engineering at Savitribai Phule Pune University for past 3 years. He has written a book on Big Data Mining -Application Perspective ISBN: 978-81-203-5116-5.