I.J. Information Technology and Computer Science, **2015, 07, 66-76**
Published Online June 2015 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2015.07.08

Issues and Challenges of User Intent Discovery (UID) during Web Search

Wael K. Hanna
Computers & Information Faculty / Information Systems Dept, Mansoura, 0000, Egypt
Email: wael_karam1@yahoo.com

Aziza S. Aseem, M. B. Senousy
Computers & Information Faculty / Information Systems Dept, Mansoura, 0000, Egypt
Sadat Academy for Management Sciences / Computer and Information Systems Dept, Cairo, 0000, Egypt
Email: dr_aziza2@hotmail.com, badr_senousy_arcoit@yahoo.com

Abstract— There is a need to a small set of words –known as a query– to searching for information. Despite the existence gap between a user's information need and the way in which such need is represented. Information retrieval system should be able to analyze a given query and present the appropriate web resources that best meet the user's needs. In order to improve the quality of web search results, while increasing the user's satisfaction, this paper presents the current work to identify user's intent sources and how to understand the user behavior and how to discover the users' intentions during the web search. This paper also discusses the social network analysis and the web queries analysis. The objective of this paper is to present the challenges and new research trends in understanding the user behavior and discovering the user intent to improve the quality of search engine results and to search the web quickly and thoroughly.

Index Terms — Query, Information Retrieval, Web Search, Social Networks, User Behavior and User Intent.

I. Introduction

Over the last two decades, the World Wide Web (Web) has been continuously growing. This rapid growth of the Web has resulted in an exponential growth of the data and information that can be found online. Thus, nowadays, the Web users have access to more information and more resources than they ever had. On the other hand, the capacity for producing information exceeds the human capacity for processing it. The need to examine large quantities of information in a limited period of time can cause the phenomenon known as information overload. Information retrieval face the following major factors that lead to information overload: excessive volume of information, difficulty or impossibility of processing it, irrelevance or non-importance of most of it, lack of time to understand it, and multiple sources containing the same information. [1]

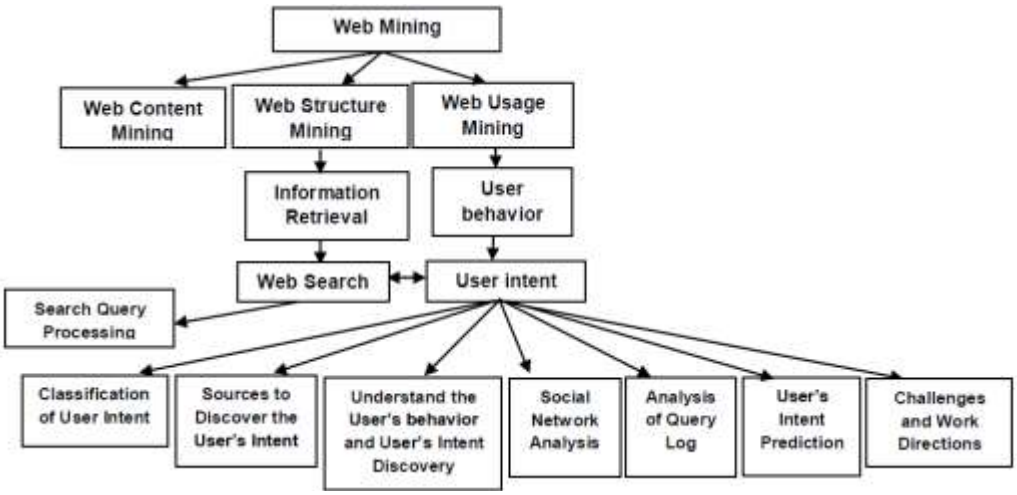

Fig. 1. Research hierarchy in user query intent in web search

The general goal of an information retrieval system is to retrieve relevant documents in order to give answer to a query. Sometimes the user is able to specify precisely what information is required for the resolution of the problem, but the common situation is that the user does not know. In order to really help the user, there is a need for web search engines to understand the kind of information a user is looking for. A small set of words – known as a query– used to searching for information. Despite the existence of this gap between a user's information need and the way in which such need is represented. The information Retrieval system should

analyze a given query and present the appropriate web resources that best meet the user's needs In order to improve the quality of results, while increasing the user's satisfaction [2], there is the need to understand and discover the users' intentions during the web search. This paper organization follows: Section two presents an introduction of web mining. Section three presents a background of information retrieval and web search and search engines. Section four presents the web query intent in web search. And finally section five contains the challenges that face user intent discovery and work directions in that area. Future work directions includes: search personalization, analysis of SN, analysis of web queries, search diversification, mobile applications & MSN, classification of web queries, new models to determine the user's intent and semantics of web queries.

Web Characteristics: There is a huge amount of linked data/information, (Data of all types) exist on the Web, Information on the Web is heterogeneous, noisy, and The Web is dynamic and virtual society. [3]

II. Related Work

This work is a survey about previous work of understanding the user intent during the web search. The objective of this paper is to present the challenges and new research trends in understanding the user behavior and discovering the user intent to improve the quality of search engine results and to search the web quickly and thoroughly.

Cristina and Ricardo [1] analyzed the impact of the query intent in the search behavior of the users. They proposed a method to identify automatically the intent behind user queries. Our paper presents current works and new challenges of classification of web queries that help search engine to understand the user intent.

Liliana and Ricardo [2] identified user's intent from a Web search engine's query log. Our paper presents current works and new challenges of analysis of query logs, query semantics and user's intent models to determine the user's intent.

Bing [3] had analyzed in his book the social network and web usage mining. Our paper presents social network analysis current works and new challenges to understand the user behavior and discover his /her intent during web search.

Alvin Chin and Daqing Zhang [4] presented mobile social networking, that is, connect with people to create social networks directly through the phone. Our paper presents current works and new challenges of mobile social networking to understand the user behavior and discover his /her intent during web search.

Dirk [5] discussed the web search engine research in different areas: Analysis of Web Search Statistics and Diversity-Aware Search. Our paper presents current works and new challenges of search personalization and search diversification to improve the search results and understand what user search for.

III. Web Mining

Web mining aims to discover useful information or knowledge from the Web. Web mining can be categorized into: 1) Web structure mining: Web structure mining discovers useful knowledge from hyperlinks. For example, discover important Web pages, which is a key technology in search engines. Discover communities of users share common interests. Web content mining: 2) Web content mining extracts or mines useful information or knowledge from Web page contents. For example, classify and cluster Web pages according to their topics and mining customer posting's to discover consumer opinions. 3) Web usage mining: The discovery of user access patterns from web usage logs, which record every click made by user. To understanding of how users behave on the Web. In particular, it is more focused on discovering the intention of the user's are when searching for information on the Web. [3]

VI. Information Retrieval and Web Search

Web information-retrieval systems appear as bridges between the users and the amount of data and information contained in the Web. The main objective of any information-retrieval system (including Web-search systems), is to satisfy the needs of the users. Some of the most popular goals that are used to describe information-retrieval systems are the following: "The goal of an information retrieval system is to locate relevant documents in response to a user's query" [1].

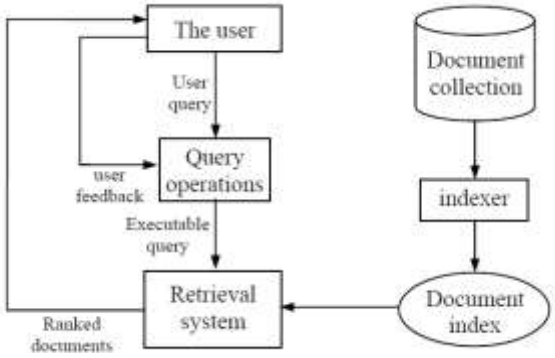

Fig. 2. A general IR system architecture [3]

Technically, IR studies the acquisition, organization, storage, retrieval, and distribution of information. Historically, IR is about document retrieval. The user with an information need issues a query to the retrieval system through the query operations module. The retrieval module uses the document index to retrieve those documents that contain some query terms, compute relevance scores for them, and then rank the retrieved documents according to the scores. The ranked documents are then presented to the user. [3]

A. Web Search

Search engine starts with the crawling of pages then parsed, indexed, and stored. At the query time, the index is used for efficient retrieval [3]

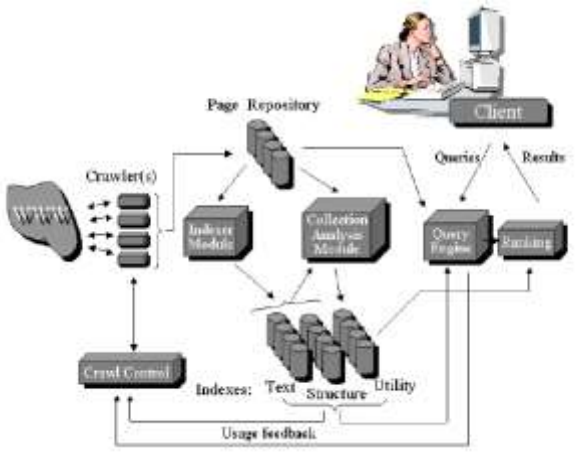

Fig. 3. General search engine architecture [6]

Crawlers are small programs follow links to reach different pages. The programs are given a starting set of URLs, whose pages they retrieve from the Web. The crawlers extract URLs appearing in the retrieved pages, and give this information to the crawler control module to determine what links to visit next. The crawlers also pass the retrieved pages into a page repository. [7]. Parsing: to parse the input HTML page to produce a stream of terms to be indexed. Indexing: search engine may build multiple inverted indices. For example, a small inverted index may be constructed based on the terms appeared in titles and anchor texts alone. A full index is then built based on all the text on each page. In searching, the algorithm may search in the small index first and then the full index. Searching and ranking: pre-processing, stopword removal and stemming. Finding pages that contain the query terms in the inverted index; ranking the pages and returning them to the user. [3]

B. Search Query Processing

To analyze the search query and deliver appropriate results, search applications rely on the "'query and result processing" engine. Queries from the user come into the query processing and transformation subsystem. This framework takes the original query, analyzes it, transforms it with and then sends the query to the search engine. The node in the search matrix that receives the query performs its retrieval operation and returns its results to the results-processing subsystem. The raw results are passed to the results-processing subsystem (which performs duplicate removal), results merging (from different search nodes), sorting, rank ordering. [7]

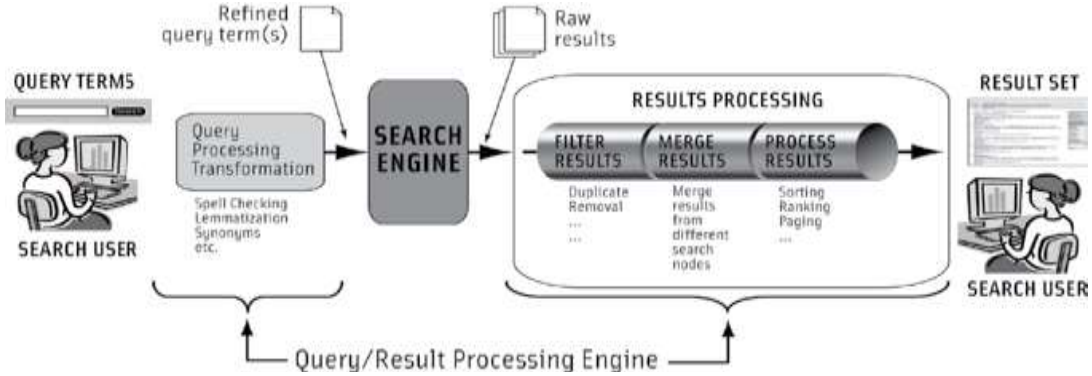

Fig. 4. Query and result processing engine [6]

V. User Query Intent in Web Search

Throughout the history of information search systems, there has been a great effort to establish the aim that leads a user to perform an information search. topics in this area range from defining the information needs of a user; passing through determine the behavior of users of Web search engines; up to establish the intention behind user queries submitted to a Web search engine. [2]

A. User's Behavior

Jansen has defined behavior as the action or specific goal–driven event with some purpose other than the specific action that is observable. Belkin pointed out that there is a fundamental importance of the users' goals, the tasks associated with those goals, with their behaviors, and the intentions underlying the behaviors, and the way they "substantially" affect their judgments of usefulness of the information objects. [2]

B. User's Intent

Broder defined intent as the need behind a query. However this definition cannot be leveraged for designing effective search engines that are aware of the mental model of the user. [8]. According to Jansen et al. the user intent is the expression of an affective, cognitive, or situational goal in an interaction with a Web search engine. Rather than the goal itself, user intent is concerned with how the goal is expressed because the expression determines what type of resource the user desires in order to address this underlying need. [2]

C. Classification of User Intent

T-he first classification Broder proposed: Navigational Informational and transactional. This taxonomy was further extended by Rose & Levinson. Refined and expanded the transactional query class with a more encompassing resource query class to include viewing, downloading and obtaining resources available on the Web. The facets as defined by Nguyen and Kan's were: Ambiguity, Authority Sensitivity, Temporal Sensitivity and Spatial Sensitivity. [2]

Kang and Kim implemented automatic classification of the queries. They classified queries into informational and navigational categories. Lee, Liu and Cho automatically classified queries into informational and navigational categories. The automatic classification based on: anchor link distribution and user click behavior. Liu et al. automatically classified navigational and transactional queries. For the automatic classification of queries used a decision tree classification algorithm and two features derived from click through data. Jansen, Booth and Spink present an automatic classification method based on attributes of each type of query, using a decision tree approach. Mendoza and Zamora automatically classify informational, navigational and transactional queries through support vector machines. Herrera et al. also automatically classify informational, navigational and transactional queries. [1]

Dirk, Jessica and Sonja [9] tested the reliability of query intent derived from queries, by a user or by a juror. It reported the findings of three studies: First, it was conducted a large-scale classification study using a crowd sourcing approach. Then, it was used click-through data from a search engine log and validated the judgments given by the jurors from the crowd sourcing study. Finally, it has conducted an online survey.

Alejandro and Gunter [10] semantically classified question-like search queries in the context yielded by preceding search queries in the same user session.

Pengjie, et al. [11] proposed a time based query classification approach to understand user's temporal intent automatically. First they analyzed the shared features of queries' temporal intent distributions. Then, they presented query taxonomy, based on temporal intents. Finally, for a new given query, they proposed a machine learning method to decide its class in terms of its search frequency over time recorded in query logs.

D.How to Understand the User's Behavior and Discover the User's Intent

The first works were based on statistical analysis of user's queries. This work allowed establishing patterns of user behavior, such as the length of queries. Since this kind of analysis is not enough to describe the user, it has emerged the idea to characterize the user's behavior in a way that describes the characteristics of the queries as well as the user's needs. [2]

Liliana and Ricardo [2] identified user's intent from a Web search engine's query log. That the user's intent can be: Informational, Not–Informational and Ambiguous. And also analyzed a set of topical categories, in which user's intent may be classified. To identify the user's intent First, a manual classification of a set of queries in order to reduce the ambiguity of some queries then, automatic identification of the user's intent. In this work, eleven out of the eighteen categories proposed were recognized by the unsupervised learning model. Finally, the author found two new and well defined categories. These categories are: cars and law. Also they introduced, analyzed and characterized a wide range of factors that may be useful for user's intent identification when searching on the Web. These dimensions/facets are: genre, topic, task, objective, specificity, scope, authority sensitivity, spatial sensitivity, and time sensitivity. Also they presented an efficient algorithm for inferring the intent of the Web search query, described in terms of multiple facets. There is scope to further improve the quality of classification by allowing powerful capabilities of WordNet to be leveraged for query intent classification.

Debora, Pinar and Sunil [8] conducted a preliminary experimental study in which users agreed to be recorded during their search activity to gathering some insights about intent.

Asli, Dilek and Gokhan [12] approached intent detection as a two-stage semi-supervised learning problem, which utilizes a large number of unlabeled queries collected from internet search engine click logs. It was captured the underlined structure of the user queries using a bayesian latent feature model. It was propagated this structure onto the unlabeled queries to obtain quality training data via a graph summarization algorithm.

Cristina and Mari [13] presented an eye-tracking study that analyzes the browsing behavior of users in the result page of search engines regarding the underlying intent of the query (informational, navigational and transactional). It was studied a diverse set of variables that influence the gaze of users: type of the search result (organic and sponsored), areas of interest and ranking position of search result. It was found that organic results are the main focus of attention for all the intent; apart from transactional queries.

Chapelle [14] et al. studied the problem of web search result diversification in the case where intent based relevance scores are available. A diversified search result will satisfy the information need of users who may have different intent.

Botao, et al. [15] observed user clicks cannot be completely explained by relevance and position bias. Users with different search intent may submit the same query to the search engine but expect different search results. Thus, there might be a bias between user search intent and the query formulated by the user, which can lead to the diversity in user clicks. It was proposed a new intent hypothesis is used to characterize the bias between the user search intent and the query in each search session.

Yuchen, et al. [16] considered the sequence of queries and their clicks in a search session as a task and proposes a task-centric click model (TCM). TCM characterizes user behavior related to a task as a collective whole. Specifically, it identified and considered two new biases in TCM as the basis for user modeling.

Nicolaas and Filip [17] presented a personalization approach that builds a user interest profile using users' complete browsing behavior, and then uses this model to rerank web results.

Giorgos and Timos [18] proposed query-centric. It examines the search behaviors/intent induced by queries and groups together queries with similar such behaviors, forming search behavior clusters. Specifically, it has exploited user feedback in terms of click data to cluster the queries. Each cluster is finally represented by a single ranking model that captures the contained intent

expressed by users. Once new queries are issued, these are mapped to the clustering and the retrieval process diversifies possible intent by combining relevant ranking functions.

User queries to the web tend to have more than one interpretation or intent due to their ambiguity and other characteristics. Chieh-Jen et al. [19] mined the subtopics of a query either indirectly or directly to diversify the search results. For the indirect subtopic mining approach, clustering the retrieval results and summarizing the content of clusters is investigated. In addition, labeling topic categories and concept tags on returned document is explored. For the direct subtopic mining approach, several external resources, such as Wikipedia, search query logs are consulted.

Vincenzo, Massimiliano, Giuseppe, and Luigi [20] on the analysis of user interactions with Search Engine Result Pages (SERPs) resulting from a web query, but most methods ignore the behavior of the user during the exploration of web pages associated to the links of the SERP s/he decides to visit. It was proposed a novel model that analyzes user interactions on such pages, in addition to the information considered by other mentioned approaches. In particular, captured user interactions are translated into features that are part of the input of a classification algorithm aiming to determine user informational, navigational, and transactional intent.

Junjun et al. [21] presented the intent mining system developed, which is capable of understanding English and Chinese query respectively, with four types of context: query, knowledge base, search results and user behavior statistics.

Aymeric, Min, Yiqun and Shaoping [22] aimed to improve the performance of search results diversification by generating an intent subtopics list with fusion of multiple resources. By thinking that to collect a large

panel of intent subtopics, it should consider as well a wide range of resources from which to extract. Such as external resources (Wikipedia, Google Keywords Generator), anchor texts, page snippets and more. It selected resources to cover both information seeker and information provider aspects.

Kerstin [5] provided an overview on diversity in web search. The reflection of a result set's coverage of multiple interpretations and intents of a query. Diversification approaches range from an adapted ranking in a way that the top results are diverse by means of similarity measures or diversity scores to a comprehensive diversity analysis which determines topics and classifies text according to opinions.

Rodrygo [23] argued that an ambiguous query should be seen as representing not one, but multiple information needs. Based upon this premise, we propose xQuAD—Explicit Query Aspect Diversification.

Yury et al. [24] studied the problem of short-term personalization specially the set of initial queries of search sessions. These, with the lack of contextual information, are known to be the most challenging for short-term personalization. They applied a widespread frame- work for personalization of search results based on the re ranking approach and evaluate our methods on the large scale data.

Jinyun et al. [25] described a characterization and evaluation of the use of cohort modeling to enhance search personalization. They experiment with three pre-defined cohorts—topic, location, and top-level domain preference—independently and in combination, and also evaluate methods to learn cohorts dynamically**.**

Harshit, Sungin and Hong [26] proposed two methods that use Singular Value Decomposition (SVD) to build a Clustered User Interest Profile (CUIP), for each user, from the tags annotated by a community of users to web resources of interest. A CUIP consists of clusters of 30 semantically or syntactically related tags, each cluster identifying a topic of the user's interest. The matching cluster, to the given user's query, aids in the disambiguation of user search need and assists the search engine to generate a set of personalized search results.

E. Social Network Analysis

Social network is the study of social entities, and their interactions and relationships. The interactions and relationships can be represented by a network or graph, where each vertex represents an actor and each link represents a relationship. From the network it can study the properties of its structure, and the role, position and prestige of each social actor. Social network analysis is useful for the Web because the Web is essentially virtual society, and a virtual social network, where each page can be regarded as a social actor and each hyperlink as a relationship. [3]

Most mobile phones include various sensors. Classification models can exploit such data to allow understanding actions and environment. Nicholas, Ye, Hong, Andrew, Tanzeem and Shane. [27] Presented the Cooperative Communities (CoCo) Framework is a new approach to personalizing classification models by leveraging social networks. The CoCo framework significantly lowers the amount of training data required from each user by sharing training data and classification models within social networks

Nicholas [28] said smart phone sensing largely oblivious to the effects of social networks and community dynamics. How might smart phone sensing systems change if they could see more than isolated individuals? What if sensing systems could not only understand these social effects, but could leverage them in their day-to-day operations — as they collect and interpret mobile sensor data?

Ming, Juanzi, Lei, and Hai-Tao [29] presented a systematic method named personalized diversity search based on user's social relationships (PDSSR), this method is a combination of personalization and diversification, which enables computer better understand user's search intent and interests, consequently returns a personalized and reduced diversified result set.

Amruta, Priyanka, Trupti, Rajeshwary and Reena [30] have solved that difficulty for the user to get the exact search results according to his preferences. The user's information will be extracted from the social networking sites like Facebook. The search keywords given by user

will be input to the NUTCH search engine. The results returned by NUTCH search engine will be further refined using our own Profile Biasing Algorithm.

Omair et al. [31] designed approach of finding the preferences of users from the relevant parts of the user's social network and community. (1) Activities of users in their SN, and (2) relevant information from user's SN, based on their proposed trust and relevance matrices.

Bin et al. [32] showed how user demographic traits such as age and gender, and even political and religious views can be efficiency and accurately inferred based on their search query histories.

Tommy et al. . [33] Used two networks, one consisting of information in web pages and the other of personal data shared on social media web sites to analyze how social media tunnels the flow of information from person to person and how to use the structure of the social network to rank, deliver, and organize information specifically for each individual user.

Martin [4] focused on social behavior in mobile social networks: first discussed different aspects of mobile social networks. Then, he summarized recent real-world analysis results, especially focusing on links between individuals, characterization of their roles, and dynamics of communities in MSN.

Daqing et al. [4] extended the definition of mobile social networks by classifying MSNs into four categories, and define two important terms, e.g., personal context and community context. Then they presented the context model and the related taxonomy of personal context and community context.

F. Analysis of Web Queries

Search query logs are files that record queries submitted by users along with the results returned by the

search engine and the results that have been clicked by the user. The formulation of a query is very important since it must -convey the exact need of the user, meaning that the words in the query should match all and only the documents being sought. Search logs of users' queries suffered from two limitations. First, queries tend to be short to correctly convey the user's intent. Second, it is difficult for users to express their information needs in such a way that it could describe the documents that they are seeking. This gap between queries and document contents is due to many reasons, including the ambiguity of some terms that have multiple meanings, as well as the existence of different words that possess the same meaning. Thus it has become imperative to go beyond the single query submitted by the user to discover the true intention of the user [3]

Analyzes of query logs from Excite, done by Jansen et al. and Spink et al., as well as the analysis done by Silverstein et al. using an AltaVista query log, reported various statistics on Web queries. Baeza–Yates showed that users from TodoCL search engine use 1.05 words per query. Spink & Jansen presented analysis of different query logs from Excite. This analysis taking into account characteristics that are exogenous to the users, such as the year query was submitted, and the geographic location the user submitted the query. A more comprehensive study was made by Spink et al. In this work, the authors presented how the search topics shifted from year 1997 to 2001. Jansen et al. conducted a comparison of nine search engines transaction logs from year 1997 to 2002. For this comparison the authors considered features such as: the sessions, queries and the results pages [2]

Baeza-Yates et al, Beeferman et al, Cao et al, Huang et al, Wen et al, Zhang et al, and Zhang et al. QLM efforts to improve user queries by suggesting or recommending modified queries, in order to retrieve more relevant documents corresponding to the user intent. Cui et al. developed a query expansion technique based mainly on the user's clickthrough data. Wen et al. mined query clusters to identify Frequently Asked Queries (FAQ). FAQs are then used to map the user's new query to suggest a new query or even to return previous verified answers. Ntoulas et al. Presented enhanced k means method was used to cluster the query logs of the Internet yellow page services for query expansion. Joachims et al. and Flake et al. supported vector machines (SVM) classification and anchor text mining were used for query modification and for query refinement, respectively**.** Fonseca et al. discovered related queries by mining association rules. Chien and Immorlica discovered semantically related search queries based on their temporal correlation. A new measure of the temporal correlation of two queries based on the correlation of their frequency functions. Davison et al. discovered queries that are related to the queries leading to a specific Web site being highly ranked. Later efforts such as in Baeza-Yates et al., Song et al. and Castillo et al. relied on user feedback in query log mining to detect polysemous queries. most early techniques to detect web and host-spam were based on content analysis or link analysis in the work of Castillo et al, Fetterly et al, Gyöngyi et al, Ntoulas et al and Wu et al. Usage data had recently started serving as a vital source of information for spam detection. For example, in Liu et al, browse logs from a tool-bar used to detect spam pages. In Castillo et al, syntactic and semantic features from the query click-graphs were extracted to detect those "query-attracting" hosts to improve spam detection. [3]

Fabrizio [34] showed the foundation of query mining and analyzing the basic algorithms and techniques that are used to extract useful knowledge from this infinite source of information. They showed search applications may benefit from this kind of analysis by analyzing popular applications of query log mining and their influence on the user experience.

Huizhon et al. [35] proposed the use of click patterns to capture the relationship among clicks on search results by treating the set of clicks made by a user as a single unit. They aggregated click patterns together using a hierarchical clustering algorithm to discover the common click patterns. By using click patterns as an empirical representation of user intent, it is able to create a rich representation of mixtures of multiple navigational and informational intents**.** They analyzed real search logs and demonstrate that such complex mixtures of intents can be identified using click patterns.

Wiestaw presented practical implications of the search statistics analysis (analyzed source of prognostic information, starting from their content and scope, their processing and applications, and concluding with usage in a software-based intelligent framework.) to observe, estimate and predict various processes using wide, precise and accurate behavior observations. It was presented potential areas that would benefit from the analysis of queries statistics. Moreover, it introduced 'WebPerceiver', an intelligent platform, built to make the analysis and usage of search trends easier and more generally available to a wide audience, including non-skilled users. [5]

Claudio et al. [36] proposed a two-step methodology for discovering tasks that users try to perform through search engines. First, identified user tasks from individual user sessions stored in search engine query logs. Second, discovered collective tasks by aggregating similar user tasks, possibly performed by distinct users.

Maksims [37] used the historical search logs to personalize top-N document rankings for a set of test users. They used over 100 features extracted from user- and query-depended contexts to train neural net and tree-based learning-to-rank and regression models.

VI. Challenges and Work Directions

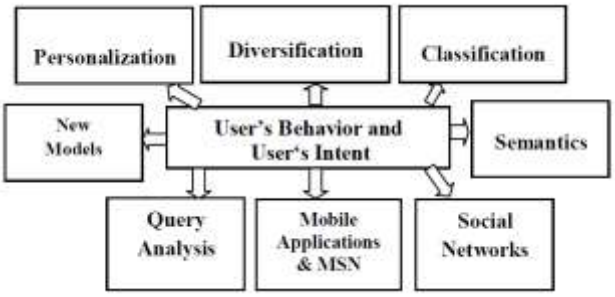

Fig. 5. Framework for new trends in UID

A. Personalization

To understand and discover the user's intent, it needs to Focusing on more than simple information about the user. Also take into account an entire spectrum of information about the user; e.g. past actions, preferences. [8]. Demographic information like age, or technological skills and information retrieval skills, as well as satisfaction stated by the participants will be taken into account in a following research to understand and discover the user's intent. [13]

The deep understanding of users' search intents when they are browsing can help extend the functionality of search engines. There are a number of extensions of using browsing history to Personalization. First, the set of parameters can still be expanded: (1) learning the parameter weights (2) using other fields, such as headings in HTML, and learning the weights for each field Also, temporal information could be incorporated: (1) investigating how much browsing history should be used (2) whether decaying the weight of older items is beneficial and (3) study how page visit duration can be usefully incorporated into the personalization algorithm. Additionally, a browser add-on has access to other behavioral information, such as time spent on a page. Similarly, it could also make use of more personal data. [17]

It will be interesting to estimate the range of applicability of filtering methods and evaluate them for other re-ranking algorithms. That is, prediction of the appearance of information need during the browsing session and detection of the dissatisfaction with the current browsing session. Develop a more complicated approach to predict the emergence of an information need. [24]

Enhancing cohort construction and cohort behavior modeling, for example, leveraging other sources of data beyond query-click logs (e.g., browsing signals, social network information) for cohort construction, and considering relationships between cohort members (e.g., group dynamics) for cohort behavior modeling. Another direction is investigating generalized cohort models. [25]

B. Social Networks

With the continued evolution of the Web and the creation of social networks new data are emerging. The study of this kind data offers a new perspective to analyze the user's intent. This information may help to visualize emerging trends of users needs. [2]. Also the social network information as the Twitter tweets starts to be shown with the results. These new features affect the way to understand and discover the user's intent, and their analysis is useful in order to improve the search engine. [13]

Systematic method named personalized diversity search based on user's social relationships (PDSSR), this method is a combination of personalization and diversification, which enables computer better understand and discover user's search intent. In the future, it would like research the influence of changing user interests to the search results. [29]

Capture other types of user traits, such as personality, intelligence, happiness, or interests and measuring the applications of those inferred traits in personalization, re ranking and monetization of the search results. [32]

The potential interaction between the searcher and the sharer is valuable because the influence of the sharer on the searcher is stronger than the influence coming from the authorities detected by HITS and PageRank in many non-technical and social situations but not for all. This feature could be implemented in search engines where pages returned to a given query are re-ranked via social networks if there are pages shared among friends or other associates of the searcher that are related to the query. [33]

Instead of refining search results based on only user's profile it can be more socialized by profiles of friends of user, family and friends of friends. [30]

The changing of user interests has influence on the user's intent. A change in user interests can be seasonal or permanent. Some changes are gradual and some are impulsive or drastic. To address these variations, a

careful balance much be drawn between long term and current user interests in user's Interest Profile. [26]

C. Analysis of web queries

Because of the gap between queries and search results, it has become imperative to go beyond the single query to discover the intention of the user. [3].

Using eye tracking some useful feedback information about the user behavior, not previously available with clickthrough information, can be envisioned. Queries can be viewed as signals in the domain of time. In each time unit record the occurrences of the query. This would result in a sort of signal to which standard temporal series. The techniques allow for the discovering of peculiar query features such as being periodic, or bursty. It can use time series to predict user behavior (i) why users issue queries and (ii) how users react to news spreading. [34]

Studying how individual user search behaviors can be exploited, and integrating query-centric method with user-centric and content-centric. [18]

Click patterns represent a significant advance as a representation of user intent, to impact a broad set of information- retrieval technologies, from the ranking to the presentation of results. [35]

It is important to analyze the evolution of the user's intent along the time. The data delivered by search engines that describe the statistics of their usage are in form of time series. There are many algorithms to deal with that form of information presentation, yet this particular situation will stimulate enhancement of time series analysis. Another stimulating factor for this area will be increasing demand for informational quality that requires filtering or removing noise without filtering out relevant information. WebPerceiver is currently in a prototype stage. Future development will include: creation of the additional modules, an interpretation mechanism that should make using the outcome of the application easier, supported creation of features, such as defining a set of search queries and embedding the data/text/web mining functions and prepared sets of features useful for a set of frequently analyzed decisional problems. WebPerceiver will likely to be extended by development of additional modules: intelligence module performs intelligent interpretation of the results, transforming the outcome of the analysis into a form more easily understandable form. Reporting module creates reports, Usenet module runs analysis of the Usenet content and Social sites module runs searches on a social sites. [5]

As future work, exploiting the collective tasks mined from the query log to build a model for representing the task-by-task search behavior of users. It used in query suggestion. [36]

Exploring contexts based on similar queries/users. Both user and query similarities can be readily inferred from the search logs using statistics like issued query overlap for users and document/domain overlap for queries. These contexts can be particularly useful for personalization of long tail queries that occur very infrequently in the data and do not have enough preference data. [37]

D. Diversification

Diversification of the search results or in search queries is inadequate to compensate for the search engine's lack of knowledge of what is in the user's mind. A general diversification scheme to truly cover all possible alternatives [8]

There are other sources to investigate the estimation approach to combine both personalization and diversification components such as of social profiles and other sources of user profile information (query and browsing history). And document classification (clustering). [38]

The Future directions include how to integrate other knowledge resources into the diversification models further, such as social information and query logs, and how to diversify Web search results with different languages. And how to localize the diversified models to meet users' needs and users' intents from different areas or countries. Detection of duplicate subtopics is expected to refine the subtopic mining performance and then enhance the performance of search result diversification. [19]

In a future work, it would like to combine sub-intent extraction, clustering and ranking data with document ranking and compare the results with some search-engines results. [22]

A future research focus will be on making the various dimensions of diversity accessible in a search result to understand and discover of user's intent. Another future challenge is to evaluate the diversity methods. Diversity analysis and result diversification are also relevant for images, and videos. It remains an interesting research issue to develop web search engines that are able to do both, diversification of text and images. [5]

A promising direction towards a pure implicit sub-query generation is a supervised approach aimed at learning the characteristics of effective sub-queries given only the top retrieved documents. Augment the user's profile by leveraging the preferences of similar users that issued the same query. For instance, such a group-based personalized diversification could be performed by exploiting the user's social circle. [23]

E. Mobile Applications & Mobile Social Networks

Today, most mobile phones include various sensors, such as GPS. Classification models can exploit such data to understand our actions and environment .Not just individual users, but the communities in which they live, can be leveraged to better model human behavior. SN can be exploited and incorporate of users within sensing systems. Hybrid sensing systems can intelligently exploit user communities in a variety of ways will overcome many of the obstacles to human modeling that currently prevent widespread usage of mobile sensing during everyday life. [27].

To progress toward general-purpose community-aware smart phone sensing systems, there is a need to more sophisticated community models. Design and architecture

of large mobile sensing systems, which would migrate from existing examples of community-aware sensing that fail to generalize beyond single narrow domains. [28]

The data quality in Mobile Social Network MSN ranging from authorized to inaccurate and even fake ones. When analyzing human behaviors from raw sensor data, it is better to train different classifiers that work in different contexts. Trust and abnormal data detection methods should be developed to ensure the trustworthiness and quality of the collected data. Raw data from different sensor sources need to be transformed to the same metrics and represented by a shared vocabulary/ontology to facilitate the learning and inference process. Data from independent sensing sources should (1) be associated, integrated, and fused to infer high-level contexts and (2) be cross-checked to allow trustworthy information inference. Better understand data produced by mobile social networks, visualization techniques and tools should be developed. In addition to the privacy issues faced by traditional online social networks, sharing and revealing personal data in MSNs are exposed to extra privacy issues that are unique to mobile environments. The problem becomes more important in short-term communities: the lack of centralized control and the anonymous-participation nature pose additional security challenges. Community information is more difficult to manage due to its complex nature. It should study the social science and domain knowledge to provide effective tools for community management. Challenges include how to extract community preferences, how to mine the underlying structure of MSNs. There is also a need to pay attention to negative social features to ensure data sharing in delay-tolerant MSNs. [4]

F. Classification of web queries

Further research should focus on improving the reliability of query classification. Classifying large data sets that can be used as baseline sets for automatic classification and bearing the following recommendations in mind (user's intent): (1) Use multiple jurors and derive multiple query intent from the data. Weight these judgments and use probabilities for the intent of each query. As jurors might disagree on the query intent, it may be useful to further ask for the reasons for such disagreement. (2) Use expert jurors and give them clear instructions and the possibility to raise a query about the classification task, as well. (3) The questionnaire or the instructions that the jurors use to classify the queries should be very detailed and perhaps also contain "traps" to detect decisions which were not made properly [9]

In future work, there is a need to explore more features for temporal intent based query classification. Also explore the application of temporal intent. Especially, Study how the temporal intent can be used to construct a page ranking model to improve information retrieval performance. [11].

G. New Models to determine the user's intent

There is a continuous need to explore and apply new machine learning models to automatically determine the user's intent. With reliable models, the search engines can aid the search process for example by adapting different ranking functions to give more accurate document positioning, adapt the number of answers and user interfaces. [2]

How to investigate more user personalized biases and consider task-related biases are a promising for designing new click models. [16]

Besides user clicks, to understand the user behavior other useful information can be derived from in click through logs, such as the user's history of input queries and visited pages. This information is related to the user's current search intent and used to identify the search intent. [15]

Incorporate some supervision to the latent factor analysis in order to collect in domain queries represented as high dimensional data. With this approach it can discover correlation between related clicked urls and user intent as prior information for latent factor analysis and build a semi-latent factor model. [12]

H. Semantics

Intent might be enriched by adding other features from the user's sessions. Also the use of semantic could give more insights to represent the user's query intent. [2]. as future work, the use of linked data for drawing additional semantic inferences assists in improving the semantic tagging. In principle, it would also be possible to build classifiers for checking as to whether or not a user input is a question-like search query, and for determining their semantic classes by some semantic database (ontology). [10]

VII. Conclusions

In order to improve the quality of search engine results, while increasing the user's satisfaction, there is the need to determine the users' intentions during the web search. The objective of this paper is to present the challenges and new research trends of understanding user's behavior and user's intent discovery to improve the quality of search engine results and to search the web quickly and thoroughly. The most interesting areas 1) to more understand and discover the user's intent, there is a need to personalizing results. Personalization should take into account information about the user and demographic information, or technological skills and their browsing history. Also enhance cohort construction and cohort behavior modeling. 2) The study of SN data offers a new perspective to analyze and understand the user's intent. Study the changing of user interests and its influence on the user's intent and web search. Capture user traits from SN can used in personalization of the search results. The potential interaction between the searcher and the sharer could be implemented in search engines. Also use social signals in web search. 3) Another area of research is the usage of click patterns and query analysis data delivered by search engines. Also WebPerceiver is currently in a prototype stage with an established list of its further enhancements and additional features and modules.

Exploring contexts based on similar queries/users can be particularly useful for web search personalization 4) Diversification covers all possible alternatives of search results to compensate for the search engine's lack of knowledge of what is in the user's mind (user's intent). Also integrating others knowledge resources as social information and query logs into the diversification models. More research in combination of diversity and personalization. A future research focus will be on making the various dimensions of diversity accessible in a search result. And evaluate the diversity methods.

Also, it considers mobile applications (mobile sensing systems) and mobile social networks to more understanding the user's behavior and discovering the user's intent. Further research should focus on improving the reliability of human classification. Classifying large data sets that can be used as baseline sets for automatic classification and bearing the following recommendations in mind (user's intent). And more future research in temporal intent based query classification.

Further research should focus on exploring and applying new machine learning models to determine the user's intent. The use of semantic could give more insights to represent the user's query intent.

REFERENCES

[1] Cristina Gonz ález-Caro, Ricardo Baeza-Yates: Supervised Identification of the User Intent of Web Search Queries, *Department of Information and Communication Technologies*, 2011.

[2] Liliana Calder ón-Benavides, Ricardo BaezaYates: Unsupervised Identification of the User's Query Intent in Web Search, *Department of Information and Communication Technologies* 2011.

[3] Bing Liu: Web Data Exploring Hyperlinks, Contents, and Usage Data , Springer-Verlag, *Berlin Heidelberg* New York, 2 ed, 2011.

[4] Martin Atzmueller: Social Behavior in Mobile Social Networks: Characterizing Links, Roles, and Communities, Daqing Zhang, Zhiyong Yu, Bin Guo, and Zhu Wang, *Mobile Social Networking Computational Social Sciences*, Exploiting Personal and Community Context in Mobile Social Networks, *Mobile Social Networking Computational Social Sciences*, Springer-Verlag, Berlin Heidelberg New York PP. 65-78, PP.109-138, 2014.

[5] Dirk Lewandowski: Web Search Engine Research. Wiestaw Pietruszkiewicz and Kerstin Denecke, The Computational Analysis of Web Search Statistics in the Intelligent Framework Supporting Decision Making and Diversity-Aware Search: New Possibilities and Challenges for Web Search, Library and Information Science, Hamburg University of Applied Sciences, Germany, PP. 79–102, PP. 139–162,2013.

[6] M.B.Senousy, Wael Karam,: A Comparative Study for Internet Search Engines and Web Crawlers", *SAMS*, Cairo, Egypt, 2011.

[7] M.B.Senousy, Wael Karam: Investigation of free open source Search Engines", *In Proceedings of the Conference on Computer Science and Software Techniques*, Czech Republic, pp.144-168, 2011.

[8] Debora Donato, Pinar Donmez, and Sunil: Toward a deeper understanding of user intent and query expressiveness: *Yahoo! Lab*, USA, 2011.

[9] Dirk Lewandowski, Jessica Drechsler, and Sonja von Mach: Deriving Query Intent from Web Search Engine Queries, *in Journal of the American Society for Information Science and Technology*, 2012.

[10] Alejandro Figueroa, Gunter Neumann: Exploiting User Search Sessions for the Semantic Categorization of Question-like Informational Search Queries, *In Proceedings of the Sixth International Joint Conference on Natural Language Processing*, Japan, PP. 902–906, 2013.

[11] Pengjie Ren, Zhumin Chen, Xiaomeng Song, Bin Li, Haopeng Yang, and Jun Ma: Understanding Temporal Intent of User Query Based on Time-Based Query Classification, *Communications in Computer and Information Science*, Vol.400. Springer-Verlag, Berlin Heidelberg New York, PP. 334-345, 2013.

[12] Asli Celikyilmaz, Dilek Hakkani Tur and Gokhan T ür: Leveraging Web Query Logs to Learn User Intent Via Bayesian Discrete Latent Variable Mode, , *In Proceedings of the of the 28th International Conference on Machine Learning*, USA, 2011.

[13] Cristina Gonz ález-Caro, Mari-Carmen Marcos: Different Users and Intent: An Eye-tracking Analysis of Web Search, *Pompeu Fabra University*, Spain, 2011.

[14] O. Chapelle, S. Ji, C. Liao, E. Velipasaoglu, L. Lai and S.-L. Wu: Intent-based Diversification of Web Search Results: Metrics and Algorithms, *Yahoo! Labs*, Microsoft Bing, CA, 2011.

[15] Botao Hu, Yuchen Zhang, Weizhu Chen, Gang Wang and Qiang Yang : Characterizing Search Intent Diversity into Click Models, *In Proceeding of the International ACM World Wide Web Conference Committee (IW3C2*), India, 2011.

[16] Yuchen Zhang, Weizhu Chen, Dong Wang, Qiang Yang,: User-click Modeling for Understanding and Predicting Search-behavior, *In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, USA, PP. 1388-1396, 2011.

[17] Nicolaas Matthijs, Filip Radlinski :Personalizing Web Search using Long Term Browsing History, *In Proceedings of the fourth ACM international conference on Web search and data mining*, USA, PP. 25-34, 2011.

[18] Giorgos Giannopoulos, Timos Sellis: Personalizing Search Results on User Intent, *NTU Athens IMIS, Greece*, 2012.

[19] Chieh-Jen Wang, Yung- Wei Lin, Ming-Feng Tsai, Hsin-Hsi Chen: Mining subtopics from different aspects for diversifying search results, *Information Retrieval*, Vol.16, No.4. Springer-Verlag, Berlin Heidelberg New York, PP. 452-483, 2012.

[20] Vincenzo Deufemia, Massimiliano Giordano, Giuseppe Polese, and Luigi Marco: Exploiting Interaction Features in User Intent Understanding, Web Technologies and Applications Lecture Notes in Computer Sciences, Vol. 7808. Springer-Verlag, Berlin Heidelberg New York, 506-517, 2013.

[21] Junjun Wang, Guoyu Tang, Yunqing Xia, Qiang Zhou, Fang Zheng, Qinan Hu, Sen Na and Yaohai Huang :Understanding the Query: THCIB and THUIS at NTCIR-10 Intent Task. *In Proceedings of the 10th NTCIR Conference on Evaluation of Information Access Technologies*, Tokyo, Japan, 2013.

[22] Aymeric Damien, Min Zhang, Yiqun Liu, and Shaoping Ma: Improve Web Search Diversification with Intent Subtopic Mining, Vol. 400. *Springer-Verlag, Berlin Heidelberg* New York, PP.322-333, 2013.

[23] Rodrygo Luis Teodoro Santos: Explicit Web Search Result Diversification, *School of Computing Science College of Science and Engineering University of Glasgow*, 2013.

[24] Yury Ustinovskiy and Pavel Serdyukov: Personalization of Web-search Using Short-term Browsing Context , *In Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, PP. 1979-1988, 2013.
[25] Jinyun Yan, Wei Chu and Ryen W. White: Cohort Modeling for Enhanced Personalized Search, *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, PP.1979-1988, 2014.
[26] Harshit Kumar , Sungin Lee, Hong-Gee Kim,: Exploiting social bookmarking services to build clustered user interest profile for personalized search, *International Journal of Information Sciences*, Vol. 281, PP. 399–417, 2014.
[27] Nicholas D.Lane, Ye Xu, Hong Lu, Andrew T. Campbell, Tanzeem Choudhury, and Shane B. Eisenman,: Exploiting Social Networks for Large-Scale Human Behavior Modeling", *Pervasive Computing, IEEE* , Vol. 10 , No. 4, PP. 45 – 53, 2011.
[28] Nicholas D. Lane :Community-Aware Smartphone Sensing Systems , *Internet Computing, IEEE*, Vol.16 No.3, PP.60 – 64, 2012.
[29] Ming Li , Juanzi Li, Lei Hou, and Hai-Tao Zheng : Personalized Diversity Search Based on User's Social Relationships, *Advanced Data Mining and Applications Lecture Notes in Computer Science*, Vol. 7713. Springer-Verlag, Berlin Heidelberg New York, PP.663-674, 2013.
[30] Amruta Mantri, Priyanka Nawale, Trupti Pardeshi, Rajeshwary Shisode, Reena Pagare : Profile Based Search Engine, *International Journal of Computer Trends and Technology*, 2013.
[31] Omair Shafiq, Tamer N. Jarada, Panagiotis Karampelas, Reda Alhajj, and Jon G. Rokne: Integrating Online Social Network Analysis in Personalized Web Search, *the Influence of Technology on Social Network Analysis and Mining, Lecture Notes in Social Networks*, Vol. 6. Springer-Verlag, Berlin Heidelberg New York, PP.589-613, 2013
[32] Bin Bi, Milad Shokouhi, Michal Kosinski and Thore Graepel, : Inferring the Demographics of Search Users , *In Proceedings of Proceedings of the 22nd international conference on World Wide Web*, ACM, PP.131-140, 2013.
[33] Tommy H. Nguyen and Boleslaw K. Szymanski : Social Ranking Techniques for the Web. *In Proceedings of IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, PP. 49-55, 2013.
[34] Fabrizio Silvestri: Mining Query Logs: Turning Search Usage Data into Knowledge, *In Journal of Foundations and Trends in Information Retriev*al, Vol.4 No.1—2, USA, PP.1-174, 2011.
[35] Huizhong Duan, Emre Kıcıman and ChengXiang Zhai, "Click Patterns: An Empirical Representation of Complex Query Intents, *UIUC Computer Science, Urbana, Microsoft Research*, USA, 2012.
[36] Claudio Lucchese, Salvatore Orlando, Raffaele Perego, Fabrizio Silvestri and Gabriele Tolomei, : Discovering Tasks from Search Engine Query Logs, *ACM Transactions on Information Systems (TOIS)*, Vol. 31 No. 3, Article No. 14, 2013.
[37] Maksims N. Volkovs : Context Models For Web Search Personalization, 2014.
[38] David Vallet and Pablo Castells: Personalized Diversification of Search Results, *Universidad Autónoma de Madrid Escuela Politécnica Superior, Departamento de Ingeniería Informática*, USA, 2012.

Authors' Profiles

Wael K. Hanna- Cairo, 5/9/1983. He is a PhD student at Computer and Information System Department, Faculty of Computer and Information, Mansoura University. He obtained his B.S. in information systems from Sadat Academy at 2004 and he got M.SC degree in information system from Sadat Academy for Management Sciences in 2011.

Now he is an assistant lecturer in Computer and Information Systems Department at Sadat Academy for Management Sciences, Cairo, Egypt. His previous research interests: Web Crawling, AI and Software Engineering. His Current research interests: Web Mining, Web Search, Information Retrieval and User Intent Discovery.

Published articles: 1) A Comparative Study for Internet Search Engines and Web Crawlers (Cairo, Egypt, Sadat Academy for management Sciences, 2011). 2) Investigation of free open source Search Engines. Information System (The 7th Annual International Conference on Computer Science & Information Systems. Athens, Greece. 13-16 June 2011.).

Hanna- personal hobbies are reading and studying.

M. B. Senousy is a professor of computer and information systems at Sadat Academy for Management Sciences, Cairo, Egypt. He has received a PhD in computer science in 1985 at George Washington University, USA.

Aziza S. Assem is a professor of computer and information systems at Faculty of Computers and Information Sciences, Mansoura, Egypt. She has received a PhD in 1996 and a MS in 1981. She has supervised master studies.

How to cite this paper: Wael K. Hanna, Aziza S. Aseem, M. B. Senousy,"Issues and Challenges of User Intent Discovery (UID) during Web Search", International Journal of Information Technology and Computer Science(IJITCS), vol.7, no.7, pp.66-76, 2015. DOI: 10.5815/ijitcs.2015.07.08

