Automated Identification of Web Queries using Search Type Patterns

Alaa Mohasseb[1], Maged El-Sayed[2] and Khaled Mahar[1]
[1]*College of Computing & Information Technology, Arab Academy for Science, Technology & Maritime Transport, Alexandria, Egypt*
[2]*Department of Information Systems & Computers, Alexandria University, Alexandria, Egypt*

Keywords: Information Retrieval, User Intent, Web Queries, Web Searching, Search Engines, Query Classification.

Abstract: The process of searching and obtaining information relevant to the information needed have become increasingly challenging. A broad range of web queries classification techniques have been proposed to help in understanding the actual intent behind a web search. In this research, we are introducing a new solution to automatically identify and classify the user's queries intent by using Search Type Patterns. Our solution takes into consideration query structure along with query terms. Experiments show that our approach has a high level of accuracy in identifying different search types.

1 INTRODUCTION

The main goal of any information retrieval system is to obtain information relevant to information needs. Search engines can better help the user to find his/her needs if they can understand the intent of the user. Identifying such intent remains very difficult; one major task in identifying the intent of the search engine users is the classification of the query type.

There are many different proposed classifications of web queries (Morrison, et al., 2001, Broder, 2002, Kellar, et al., 2006, Baeza-yates, et al., 2006, Ashkan, et al., 2009, Lewandowski, et al., 2012, Bhatia, et al., 2012). Broder's classification of web queries (Broder, 2002) is one of the most commonly used classifications. It classifies web queries to three main types: Informational queries, Navigational queries and Transactional queries.

Some researches (Choo, et al. 2000, Morrison, et al., 2001, Broder, 2002, Rose, et al., 2004, Kellar, et al., 2006) used different manual methods to classify users' queries like surveys and field studies. Other researches used automated classification techniques like supervised learning, SVM…etc. (Lee, et al., 2005, Beitzel, et al., 2005, Baeza-yates, et al., 2006, Liu, et al., 2006, Ashkan, et al., 2009, Mendoza, et al., 2009, Jansen, et al., 2010, Kathuria, et al., 2010).

One drawback of the solutions that were introduced so far is that they do not take into consideration the structure of the queries. Queries submitted to search engines are usually short and ambiguous and most of the queries might have more than one meaning, therefore using only the terms to identify search intents is not enough, two queries might have exactly the same set of terms but may reflect two totally different intents, therefore classifying web queries using the structure of the query in addition to terms and characteristics may help in making the classification of queries more accurate.

In our research, we propose a solution that automatically identifies and classifies user's queries using Search Type Patterns. Such Search Type Patterns are created from studying different web queries classification proposals and from the examination of various web logs. A Web Search Pattern is constructed from one or more terms, such terms are categorised and introduced in the form of taxonomy of search query terms.

We have developed a prototype to test the accuracy of our solution. Experimental results show that our solution accurately identified different search types.

The rest of the paper is organized as follows: Section 2 highlights the different proposed classification techniques used in web query identification. Section 3 provides detailed explanation of the extended classification of web search queries and the different type of each category. Section 4 provides a detailed description of the proposed solution. Section 5 covers experiments and results and finally Section 6 gives

Mohasseb A., El-Sayed M. and Mahar K..
Automated Identification of Web Queries using Search Type Patterns.
DOI: 10.5220/0004849402950304
In *Proceedings of the 10th International Conference on Web Information Systems and Technologies* (WEBIST-2014), pages 295-304
ISBN: 978-989-758-024-6

conclusion and future work.

2 PREVIOUS WORK

2.1 Search Types

According to (Broder, 2002) web searches could be classified according to user's intent into three categories: Navigational, Informational and Transactional. Many researches (Liu, et al., 2006, Jansen, et al., 2008, 2010, Mendoza, et al., 2009, Kathuria, et al., 2010, Hernandez, et al., 2012) have based their work on Broder's classification of user query intent. Others like (Lee, et al., 2005) used navigational and informational queries only due to lack of consensus on transactional query and to make classification task more manageable.

Rose, et al., (2004) and (Jansen, et al., 2008) extended the classification of Informational, Navigational and Transactional queries by adding level two and level three sub-categories.

Lewandowski, et al., (2012) proposed two new query intents, Commercial and Local. According to their work, the query might have a Commercial potential like the query: *"commercial offering"* or the user might search for information near his current location.

Bhatia, et al., (2012) classified queries to four classes: Ambiguous, Unambiguous but Underspecified, Information Browsing and Miscellaneous.

Calderon-Benavides, et al., (2010) and Ashkan, et al., (2009) proposed other classification of queries that classified user intent into dimensions and facets. These dimensions and facets are extracted from user's queries to help the identification of user intent when searching for information on the web like Genre, objective, specificity, scope, topic, task, authority sensitivity, spatial sensitivity and time sensitivity (Calderon-Benavides, et al., 2010).

Ashkan, et al., (2009) classified query intent into two dimensions, Commercial and Non-commercial and Navigational and Informational.

Kellar, et al., (2006) classified web informational task based on three main informational goals, Information Seeking, Information Exchange and Information Maintenance.

Baeza-yates, et al., (2006) established three categories for user search goal, Informational, Not Informational and Ambiguous. Informational query when the user's interest is to obtain information available on the web. Not Informational include specific transactions or resources like *"buy"*, *"download"*...etc. Ambiguous queries include queries that can't be identified directly because the user interest is not clear.

Morrison, et al., (2001) classified search goals into Find, Explore, Monitoring and Collect, this classification focus on three variables: the purpose of the search, the method used to find information and the contents of searched information.

2.2 Classification Methods and Techniques

Researchers have used different manual and automated classification methods and techniques to identify users query intent.

Broder, (2002) classified user's query manually by using a survey of AltaVista users as one of the methods to determine the type of queries, the survey was done online and users were selected randomly. Users were asked to describe the purpose of their search, queries that were neither Transactional nor Navigational were assumed to be Informational, the final results of the survey showed that 24.5% of the queries were Navigational, Informational queries accounted for 39% of the queries and transactional accounted for 36% of the queries. In addition Broder has analysed a random set of 1000 queries from the daily AltaVista log, queries that were neither Transactional nor Navigational were assumed to be Informational, results showed that 20% of queries were Navigational, 48% were Informational and 30% were Transactional.

Choo, et al., (2000) and Kellar, et al., (2006) used questionnaire survey for manual classification of queries and since participants in this kind of classification were low in number, the results can't be considered reliable.

In addition to the questionnaire survey (Kellar, et al., 2006) conducted one-week field study to classify data using a custom web browsing and analysed the data for only 21 participants.

Rose, et al., (2004) argued that user goals can be deduced from looking at user behaviour available to the search engine like the query itself, result clicked…etc. This approach has limitation that the goal-inferred from the query may not be the user actual goal.

Lewandowski, et al., (2012) analysed click-through data to determine Commercial and Navigational queries and used crowdsourcing approach to classify a large number of search queries.

Liu, et al., (2006) also used click-through data for query type identification. Queries were randomly

selected and manually classified by three assessors using voting to decide queries category. This work relied on decision tree algorithm and used precision and recall to test effectiveness of the query type identification.

Lee, et al., (2005) proposed two types of features, past user click behaviour and Anchor-link distribution. Results showed that the combination of these two techniques could correctly identify the goals for 90% of the queries.

Hernandez, et al., (2012) introduced a solution that automatically classifies queries using only the text included in the query, based on the feature and characteristics described by (Broder 2002, Jansen, et al., 2008, Dayong, et al., 2010). More than 1692 queries were manually classified then two machine-learning algorithms, naïve Bayes and Support Vector Machine (SVM), were used. Results showed that the two machine-learning algorithms suited more Informational and Transactional queries; results of Navigational queries were very low with naïve Bayes and null with SVM. These Results indicate that using only the content of words in the queries is not sufficient to find all user intents.

Ashkan, et al., (2009) classified 1700 queries and manually labelled the selected queries then used ads click-through and query features to determine the query intent.

Beitzel, et al., (2005) and Baeza-Yates, et al., (2006) used supervised learning to determine query intents. In addition to supervised learning (Baeza-Yates, et al., 2006) applied unsupervised learning then combined both techniques to identify user search goal.

Jansen, et al., (2008) developed a software application that automatically classified queries using web search engine log of over a million and a half queries. Results showed that more than 80% of web queries were Informational, Navigational and Transactional queries each represent about 10% of web queries. To validate their approach 400 queries from Dogpile transaction log were randomly selected and manually coded, 74% of the queries were successfully classified and the remaining 25% were vague or multi-faceted queries.

Kathuria, et al., (2010) automatically classified queries using k-means clustering, results for this technique showed that more than 75% of web queries are Informational in nature and 12% each for navigational and transactional queries.

3 BACKGROUND

3.1 Web Search Queries Classification

The following sections describe in details each of the categories we considered in our work. These categories are based on work done by (Broder, 2002, Rose, et al., 2004 and Jansen, et al., 2008).

3.1.1 Informational Searching

Informational Searching has five sub-categories:

a) Informational - Directed (I, D): the goal of this category is to learn something in particular about a certain topic, or to answer a specific question, both open and closed ended. This category has level two sub-categories:

a.1) Informational - Directed - Open (I, D, O): this category may take many forms either a question to get an answer for an open-ended question or one with unconstrained depth or to find information about two or more topics. Examples: *"why are metals shiny?"* and *"honeybee communication"*.

a.2) Informational - Directed - Closed (I, D, C): queries in this category can be a question to find one specific or unambiguous answer or to find information about one specific topic. Examples: *"capital of Brazil"* and *"what is a prime number?"*

b) Informational - Undirected (I, U): the purpose of this category is to know anything and everything about a topic, most queries in this type are related to science, medicine, history and news and celebrities (Rose, et al., 2004). Examples: *"Shawn Johnson"*, *"Vietnam war"* and *"hypertension"*.

c) Informational - List (I, L): plural query terms are a highly reliable indicator of this category (Rose, et al., 2004), the goal of this type of queries is to find a list of suggested websites or candidates or list of suggestions for further research. Examples: *"list of Disney movies"*, *"London universities"*, and *"things to do in Atlanta"*.

d) Informational - Find (I, F): the goal of this category is to find or locate something in the real world like a product or service. Most product or shopping queries have the locate goal (Rose, et al., 2004), for example: *"apple store location in New Jersey"* and *"cheap apple MacBook pro"*.

e) Informational - Advice (I, A): the goal of this category is to get ideas, suggestions, advice or instructions about something and may take many forms like a question. Examples: *"How to download iTunes"* and *"writing a book"*.

3.1.2 Navigational Searching

Navigational Searching has two sub-categories:

a) Navigational to Transactional (N, T): the URL or website user is searching for is a transactional site. Examples: *"amazon.com"* and *"ebay.com"*.

b) Navigational to Informational (N, I): the URL or website user is searching for is an informational site. Examples: *"google.com"* and *"yahoo.com"*.

3.1.3 Transactional Searching

Transactional Searching has the following sub-categories:

a) Transactional - Obtain (T, O): the goal of this type of queries is to obtain specific resource or object, not to learn some information but just to use the resource itself. This category has the following level two sub-categories:

a.1) Transactional - Obtain - Online (T, O, O): the resources of this type of queries will be obtained online, meaning that the user might search for something to just look at it on the screen. Examples: *"meatloaf recipes"* and *"Adele Songs lyrics"*.

a.2) Transactional - Obtain - Offline (T, O, F): the resources of this type of queries will be obtained offline and may require additional actions by the user, meaning that the user might search for something to print or save to use it later offline. Examples: *"Bon Jovi wallpapers"* and *"windows 7 screensavers"*.

b) Transactional - Download (T, D): the resource of this type of query is something that needs to be installed on a computer or other electronic device to be useful like finding a file to download. This category has level two sub-categories:

b.1) Transactional - Download - Free (T, D, F): the downloadable file is free. Examples: *"free online games"* and *"free mp3 downloads"*.

b.2) Transactional - Download - Not Free (T, D, N): the downloadable file is not necessarily free. Examples: *"safe haven book download"* and *"Kelly Clarkson songs download"*.

c) Transactional - Interact (T, I): this type of queries occurs when the intended result of the search is a dynamic web service, and requires further interaction with a program or a resource. Examples: *"currency converter"*, *"stock quote"*, *"buy cell phones"*, and *"weather"*.

d) Transactional - Results Page (T, R): the goal of this category is to obtain a resource that can be printed, saved, or read from the search engine results page. This category has level two sub-categories:

d.1) Transactional - Results Page - Links (T, R, L): the resources of this kind of queries appear in the title, summary, or URL of the search engine results page. Example: *"searching for title of a conference paper to locate the page numbers"*.

d.2) Transactional - Results Page - Other (T, R, O): the resources of this kind of queries does not appear on the search engine results page but somewhere else on the search engine results page. Example: *"spelling check of a certain term"*.

3.2 Characteristics of Web Search Queries

3.2.1 Informational Search Characteristics

One of the major characteristics of Informational Searching is the use of natural language phrases (Jansen, et al., 2008). Queries for such search may consist of informational terms like *"list"* and *"playlist"*…etc., question words like *"who"*, *"what"*, *"when"*…etc. Searches related to Advice, help and guidelines like *"FAQs"* or *"how to"*…etc., ideas and suggestions terms, recent information and news like *"weather"*.

Some queries consisting of multimedia like videos are considered informational like *"how-to-do"* videos. Topics related to science, medicine, history, news and celebrities are also considered informational, (Rose, et al., 2004).

3.2.2 Navigational Search Characteristics

Navigational Searching queries contain organization, business, company and universities name, domain suffixes like *".com",".org"*…etc. also prefixes such as *"www"* or *"http"* and "*web*" as the source. Some Navigational queries contain URLs or parts of URLs (Jansen, et al., 2008).

Most queries consisting of people names, including celebrities, are not considered navigational. According to (Rose, et al., 2004) a search for a celebrity such as *"Justin Timberlake"* will result in a fan or media sites, and usually the goal or objective of searching for a celebrity is not just visiting a specific site.

3.2.3 Transactional Search Characteristics

According to (Jansen, et al., 2008) queries in Transactional Searching is related to obtaining terms like *"lyrics"*, *"recipes"*, *"patterns"*…etc., download terms like, *"software"*…etc. Also queries containing

"audio", *"video"* and *"images"* are considered to be transactional.

Queries related to entertainment terms like *"pictures"*, *"games"*…etc., and e-commerce. Interact terms such as *"buy"*, *"chat"*, *"book"*, *"order"*...etc., and file extensions like *"jpeg", "zip*...etc., (Jansen, et al., 2008).

4 PROPOSED SOLUTION

Our solution mainly relies on Search Type Patterns (STPs). These patterns generalize web search queries of different types and could be used in identifying the query class and hence the user's intent. We have constructed 1182 different Search Type Patterns. Examples of these patterns are given in sections 4.1 and 4.2. Due to space limitation we couldn't give a comprehensive listing of these patterns.

Our proposed Search Type Patterns cover all categories discussed in section 3.1 above except Navigational search sub-categories and the Transactional-Results page category. The reason of excluding these categories is because it is not possible to determine the intent of the query without performing the search and monitoring the user's interaction with the result, which falls outside the scope of our work since our solution is not based on processing the search results. For example, if a user searches for: "UCLA University", he might be interested in browsing the site to know more information (Navigational-to-Informational) or to register a course (Navigational-to-Transactional).

Each Search Type Pattern (STP) is composed of a sequence of term categories (tc). STP = <tc_1, tc_2,…, tc_n>. Each term category tc_i contains a list of terms. The categorization of terms in our solution is mainly based on the seven major word classes in English: Verb, Noun, Determiner, Adjective, Adverb, Preposition and Conjunction. In addition to that we added a category for question words that contains the six main question words: How, who, when, where, what and which. We further extended this classification by adding two super-categories: Domain Suffixes and Prefixes. We also added sub-categories where a category may have one or more sub-categories.

Term sub-categorization is built in a way that enables the preservation of uniqueness of each Search Type Pattern. In other words, no two Search Type Patterns will have exactly the same sequence of term categories. Section 4.1 will discuss in details how term categorization and Search Type Patterns were constructed.

Table 1 shows detail of all term categories in our solution and Figure 1 shows the taxonomy organization of these categories.

Table 1: List of Term Categories.

Category Name	Abbreviation	Terms
Action Verb-Interact terms	AV_I	Buy, Reserve, Order…etc.
Action Verb-Locate	AV_L	Locate, Find.
Action Verb-Locate & Interact terms	AV_IL	All Locate & Interact terms.
Action Verb-Download	AV_D	Download
Action Verbs	AV	Write, create, drive…etc.
Auxiliary Verb	AuxV	Can, may, will…etc.
Linking Verbs	LV	Is, are, was…etc.
Verbs	V	All Verbs
Adjective Free	Adj_F	Free
Adjective Online	Adj_O	Online
Adjective Free & Online	Adj_OF	Free & Online
Adjective	Adj	All Adjectives
Adverb	Adv	Almost, barely, highly…etc.
Determiners	D	A, An, The…etc.
Conjunction	Conj	And, as, but…etc.
Ordinal Numbers	NN_O	1st, second, 70th…etc.
Cardinal Numbers	NN_C	1, 50, ten...etc.
Numeral Numbers	NN	All numbers
Celebrities Name	PN_C	Phil Collins, Clint Eastwood, The Beatles…etc.
Entertainment	PN_Ent	Specific name of a song, movie, game…etc.
Newspapers, Magazines, Documents, Books...etc.	PN_BDN	Specific name of a Newspapers, Magazines, Documents, Books...etc.
Events	PN_E	Cannes film festival…etc.
Celebrities, Events, Newspapers, Entertainment…etc.	PN_BCEE	All PN_C, PN_BDN, PN_Ent & PN_E
Companies Name	PN_CO	IBM, Microsoft, Intel...etc.
Geographical Areas	PN_G	London, Europe, Nile River…etc.
Places and Buildings	PN_PB	Eiffel Tower, National park…etc.

Table 1: List of Term Categories. (Cont.)

Category Name	Abbreviation	Terms
Institutions, Associations, Clubs, Parties, Foundations and Organizations	PN_IOG	Yale university, Warren middle school…etc.
Companies, Geographical Areas, Institutions, Places…etc.	PN_CGIP	All PN_CO, PN_G, PN_PB & PN_IOG
Celebrities, Entertainment, Companies…etc.	PN_BCC	All PN_BCEE & PN_CGIP
Brand Names	PN_BN	Coach, Pepsi, Gucci…etc.
Software & Applications	PN_SA	uTorrent, Photoshop, Skype…etc.
Products	PN_P	iPad, Oreo cookie…etc.
Brand, Products, Software…etc.	PN_BSP	All PN_BN, PN_P and PN_SA
Brand, Products, Entertainment, Companies…etc.	PN_BBC	All PN_BCC & PN_BSP
History and News	PN_HN	Revolutionary war, American Civil war…etc.
Religious Terms	PN_R	Christian, Muslim, God, Allah…etc.
Holidays, Days, Months	PN_HMD	Christmas, Saturday, November…etc.
Religious Terms, Holidays, Days, Months	PN_HR	All PN_R & PN_HMD
Health Terms	PN_HLT	Specific Terms related to health & medicine.
Science Terms	PN_S	Specific Terms related to Science.
Health & Science Terms	PN_HS	All PN_S & PN_HLT
Proper Noun	PN	All Proper Nouns
Database and Servers	CN_DBS	Weather, Dictionary…etc.
Advice	CN_A	Advice, ideas, instruction, suggestion, tips.
Download	CN_D	Download, Software
Entertainment	CN_Ent	Music, Movie, Sport, Picture, Game…etc.
File Type	CN_File	MP3, PDF…etc.
Informational Terms	CN_IFT	List, Playlist…etc.

Table 1: List of Term Categories. (Cont.)

Category Name	Abbreviation	Terms
Info. Terms, File & Entertainment	CN_EFI	All CN_Ent, CN_File & CN_IFT
Obtain Offline	CN_OF	Wallpapers, documents…etc.
Obtain Online	CN_OO	Lyrics, Recipes…etc.
Obtain	CN_OB	Obtain Online & Offline
File, Entertainment, Informational & Obtain Terms	CN_OBEF	All CN_EFI & CN_OB
History & News	CN_HN	History, News, War, Rumour.
Interact terms	CN_I	Translation, reservation…etc.
Locate	CN_L	Location
Site, Website, URL	CN_SWU	Site, Website, URL, Webpage.
Common Noun – Other- Singular	CN_OS	All singular common nouns
Common Noun-Other- Plural	CN_OP	All plural common nouns
Common Noun-Other	CN_O	Other Common Nouns
Common Noun	CN	All Common Nouns
Pronoun	Pron.	I, Me, You…etc.
Noun	N	All Nouns
Domain Suffix	DS	.com, .org, .us…etc.
Prefixes	DP	http, www.
Preposition	PP	For, of, about…etc.
How	QW_How	How, How far, How many, How much, How often
What	QW_What	What
When	QW_When	When
Where	QW_Where	Where
Who	QW_Who	Who
Which	QW_Which	Which
Question Words	QW	All question words

4.1 Constructing Search Type Patterns and Term Category Taxonomy

In order to construct Search Type Patterns and term categories we have used 80,000 randomly selected queries from AOL 2006 datasets. We have taken the following steps:

Step 1- parsing the 80,000 queries and automatically extracting terms in the queries.

Step 2- manually performing initial categorization for the terms.

Step 3- processing the queries and converting each query to a Query Pattern. A Query Pattern (QP) is a representation of the original query where each term is replaced by a term category from the categories that we have constructed. QP = <tc_1, tc_2, …, tc_n>. For example, the query: "*Free Wallpapers"* is converted to the Query Pattern: *<Adj_F + CN_OF>*. A Query Pattern is an intermediate step towards reaching the final refined Search Type Patterns.

Step 4- grouping similar queries according to their Query Pattern. This reduced the size of the initial dataset significantly. For example, the two queries: "*Who is Stephen Hawking"* and "*Who is Michael Phelps"* both have the same Query Pattern: *<QW_Who + LV + PN_C>*. The resulting set of Query Patterns is a much smaller representation of the original dataset.

Step 5- manually classifying each Query Pattern into one of the search types discussed in section 3.1. According to the semantics of the search types. For example, the Query Pattern: *<QW_Who + LV + PN_C>* is classified as *Informational-Directed-Closed.*

Step 6- To reduce the number of resulting patterns and to make them more generalized we performed a final step where we analysed Query Patterns in each search types separately and merged patterns that could be merged. Two Query Patterns QP_x = <tc_{x1}, tc_{x2}, …, tc_{xn}> and QP_y = <tc_{y1}, tc_{y2}, …, tc_{yn}> could be merged if for each $tc_{xi} \in QP_x$ and $tc_{yi} \in QP_y$ $tc_{xi} = tc_{yi}$ or there is a common super-category tc_{sup} for both tc_{xi} and tc_{yi}. Such super-category might already exist or it might be created, in this case we merge the two Query Patterns QP_x and QP_y in one new pattern that contains tc_{sup} instead of tc_{xi} and tc_{yi}. For example, the four Query Patterns: *<CN_L + PP + PN_PB>* (representing the query: "*Location of Eiffel Tower"*), *<CN_L + PP + PN_G>* (representing the query: "*Location of Kuwait"),* *<CN_L + PP + PN_IOG>* (representing the query: "*location of university of Florida"), and <CN_L + PP + PN_CO>* (representing the query: "*location of IBM")* are merged into the Query Pattern: *<CN_L + PP + PN_CGIP>,* since PN_PB, PN_G, PN_IOG and PN_CO term categories have the same super-category *PN_CGIP*. Note that this step has resulted in the final refined taxonomy of term categories presented in Figure 1 and Table 1.

The final set of Query Patterns after merging is called the Search Type Patterns. Note that if sub-categories that are being merged are not representing all the terms in the super-category we still use the super-category if we found that the new Query Pattern is valid for the Search Type. This helped in making our patterns covering more queries than those just being encountered in the input dataset.

As a result of applying the steps above we generated a database that contains all terms extracted from the dataset that we have used. We enriched this database by adding all possible terms in all the 7 main super-categories except the Proper Noun Category, since Proper Nouns are infinite. Note that although our solution does not require knowing all Proper Nouns, it is still capable of classifying queries that contain unrecognized Proper Nouns, as

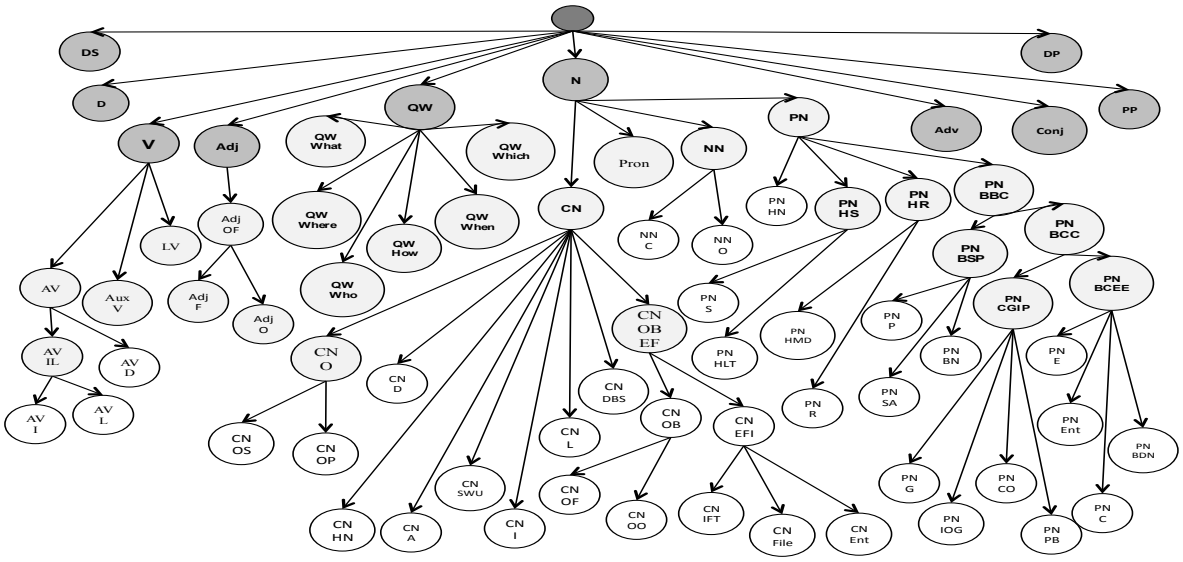

Figure 1: Terms Categorization.

we are going to illustrate in the next Section. The resulting database contains 10,440 terms classified into the classes shown in Table 1.

In addition to the term categories, we were able to identify 1182 Search Type Patterns. Table 1 and Table 2 show the distribution of these patterns by search type.

We validated our Search Type Patterns using a dataset containing 1953 queries from AOL that were manually classified and used in (Mendoza, et al., 2009).

Table 2: level 1 Search Type Patterns Distribution.

Type of search	Total
Informational	838
Transactional	336
Navigational	8

Table 3: Level 2 and Level 3 Search Type Patterns Distribution.

Type of search	Total
Informational -List	155
Informational -Find	164
Informational -Advice	121
Informational -Undirected	51
Informational -Directed -Open	113
Informational -Directed -Closed	234
Transactional -Obtain - Online	59
Transactional -Obtain -Offline	76
Transactional -Interact	28
Transactional -Download –Free	104
Transactional -Download -not Free	69

4.2 Classifying Search Engine Queries

Our solution automatically identifies and classifies user's queries by utilizing the Search Type Patterns and the term categories taxonomy presented in Figure 1. The proposed solution has three phases as shown in Figure 2:

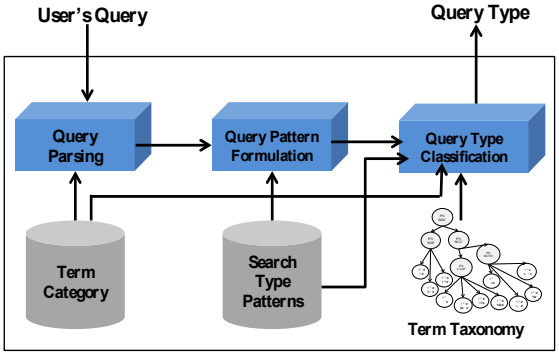

Figure 2: Proposed System Framework.

Phase 1- query parsing: this step is mainly responsible for extracting user's query terms. Unlike most other ir solutions, our solution does not destroy the query structure by removing stop-words and wh-question words. Such important query components are exploited in determining the query type. The system simply takes the user's query and parses it to facilitate the mapping of each word to the right category. For example given the two queries: query 1: *"what is the capital of romania?"* And query 2: "*list of movies by steven spielberg"* as inputs, the system extracts the following terms from query 1: *"what", "is", "the", "capital", "of", "romania"*, and extracts the following terms from query 2: "*list", "of", "movies", "by", "steven spielberg"*.

Phase 2- Query Pattern Formulation: the system converts the query to a Query Pattern by mapping terms in the query to corresponding term categories. First the system checks for compound terms (phrases) and then it processes single terms. The system maps each term to the most specific sub-category. If a term is not found in the terms database, the system assumes that the term is a Proper Noun, since Proper Nouns are infinite and we do not maintain an exhaustive list of them. After determining term category for all terms in the user query we then process consecutive terms that were identified as Proper Nouns. We convert such sequence of Proper Nouns to a single Proper Noun since no Search Type Pattern contains consecutive independent Proper Nouns.

The result of applying step 2 to query 1 is: *"What"→QW_What, "is"→LV, "the"→D, "capital"→CN_OS, "of"→PP, "Romania"→PN_G.* As a result, the Query Pattern for query 1 is: *<QW_What + LV + D + CN_OS + PP + PN_G>*. For query 2, if the terms database contains *"Steven Spielberg"*, the system will be able to identify *"Steven Spielberg"* as a phrase and to determine its type as *PN_C*, hence the system will generate this Query pattern for query 2: *<CN_IFT + PP + CN_Ent + PP + PN_C>*. If *"Steven Spielberg"* was not contained in the terms database, the system assigned *"Steven"*→PN and *"Spielberg"*→PN, since both were not identified as any other type. The system then constructs this initial Query Pattern for query 2: *<CN_IFT + PP + CN_Ent + PP + **PN** + **PN**>* then it is modified to *<CN_IFT + PP + CN_Ent + PP + **PN**>* by merging the two consecutive Proper Nouns into a single Proper Noun.

Phase 3- Query Type Classification: In this step the system attempts to match the Query Pattern generated in step 2 with the most appropriate Search

Type Patterns to determine the Query type. For some Query Patterns, like the Query Pattern of query 1, this will be straightforward. This Query Pattern matches a Search Type Pattern in the Search Type *Informational-Directed-Closed*.

For other queries, like query 2, the Query Type does not fully match any Search Type Pattern. In this case we retrieve all Search Type Patterns that partially match the Query Pattern and we use the term categories taxonomy to determine which Search Type Patterns better match the Query Pattern. For example the Query Pattern <*CN_IFT + PP + CN_Ent + PP + PN*> of query 2 partially matches the Search Type Pattern <*CN_IFT + PP + CN_Ent + PP + PN_C*> from the *Informational-List* search type. And since PN_C is a sub-category of PN, the system classifies query 2 as *Informational-List*. Note that if the Query Pattern partially maps to a single search type, we can use this as a knowledge-learning step as the system might automatically add the new ambiguous term to the term categories database. This enriches the database of the system and reduces the cases of term ambiguity and partial query type matching in the future. If the Query Pattern partially maps to multiple search types, the system classify the query to more than one search types. This is a better treatment than considering the query totally vague and discarding it, as done by other solutions. This could be used to reduce the size of search engine result as we can provide the user with a very limited number of options that would reflects his/her intention.

5 EXPERIMENTS

We developed a prototype in Java to test our proposed solution. Our prototype utilizes the 1182 different Search Type Patterns that we have constructed and also use the taxonomy of term categories shown in Figure 1 and Table 1. This taxonomy of term categories contains 10,440 different terms and types.

To test the accuracy of our solution, 10,000 queries were randomly selected from AOL 2006 dataset and tested using the system. The selected queries are different from those used in constructing the Search Query Patterns. Results of the experiment show that our solution had identified and classified 7754 of the queries. After examining the remaining unclassified 2246 queries, we found that 927 of them were not identified due to vagueness or mistakes. This make the accuracy of the classification 85.5% of the queries without mistakes.

Table 4, shows classification detail by search type. Informational queries have the highest frequency with 4245 queries then transactional queries with 2783 queries. Navigational queries have the lowest frequency with only 726 queries. Table 5 shows the breakdown of the result to sub-categories.

Our experiments show that 944 out of the 1182 different Search Type Patterns were used in classifying the 10,000 queries that were used in our experiment.

Table 4: Query Classification Results.

Type of search	Total
Informational	4245
Transactional	2783
Navigational	726

Table 5: Extended Classification Results.

Type of search	Total
Informational -List	1117
Informational -Find	875
Informational -Advice	351
Informational -Undirected	986
Informational -Directed -Open	283
Informational -Directed -Closed	633
Transactional -Obtain -Online	860
Transactional -Obtain -Offline	726
Transactional -Interact	94
Transactional -Download -Free	548
Transactional -Download -not free	555

6 CONCLUSIONS

In this research, we have introduced a framework to automatically identify and classify search engine user queries. Unlike other solutions, our solution relies on both query terms and query structure in order to determine the user intent. We have categorized search queries through introducing Search Type Patterns. Our framework consists of three main steps: (1) parsing user's query, (2) formulating Query Patterns, and (3) Classifying query type.

Experiments show that our solution has achieved high accuracy in classifying queries. As a future work we will examine and analyze more queries from different search engine datasets in order to extend the ability of our system to identify more queries. We also plan to conduct more experiments on larger datasets and compare our results to results obtained from other approaches.

REFERENCES

Ashkan, A., Clarke, C. L.,Agichtein, E., &Guo, Q., 2009. Classifying and characterizing query intent.In*Advances in Information Retrieval* (pp. 578-586).Springer Berlin Heidelberg.

Broder, A., 2002.A taxonomy of web search.In *ACM Sigir forum* (Vol. 36, No. 2, pp. 3-10). ACM.

Bhatia, S., Brunk, C., &Mitra, P., 2012. Analysis and automatic classification of web search queries for diversification requirements. *Proceedings of the American Society for Information Science and Technology*, 49(1), 1-10.

Baeza-Yates, R., Calderón-Benavides, L., & González-Caro, C., 2006.The intention behind web queries.In*String processing and information retrieval* (pp. 98-109).Springer Berlin Heidelberg.

Beitzel, S. M., Jensen, E. C., Frieder, O., Grossman, D., Lewis, D. D., Chowdhury, A., &Kolcz, A., 2005. Automatic web query classification using labeled and unlabeled training data. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 581-582). ACM.

Choo, C. W., Detlor, B., & Turnbull, D., 2000. Information seeking on the Web: An integrated model of browsing and searching. *firstmonday*, 5(2).

Calderón-Benavides, L., Gonzalez-Caro, C., & Baeza-Yates, R., 2010. Towards a deeper understanding of the user's query intent. In *SIGIR 2010 Workshop on Query Representation and Understanding* (pp. 21-24).

Hernández, D. I., Gupta, P., Rosso, P., & Rocha, M. A., 2012.Simple Model for Classifying Web Queries by User Intent.

Jansen, B. J., & Booth, D., 2010. Classifying web queries by topic and user intent. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems* (pp. 4285-4290). ACM.

Jansen, B. J., Booth, D. L., & Spink, A., 2008. Determining the informational, navigational, and transactional intent of Web queries. *Information Processing & Management*, 44(3), 1251-1266.

Kathuria, A., Jansen, B. J., Hafernik, C., &Spink, A., 2010. Classifying the user intent of web queries using k-means clustering. In *Internet Research*, 20(5), 563-581.

Kellar, M., Watters, C., & Shepherd, M., 2006.A Goal-based Classification of Web Information Tasks.*Proceedings of the American Society for Information Science and Technology*, 43(1), 1-22.

Liu, Y., Zhang, M., Ru, L., & Ma, S., 2006. Automatic query type identification based on clickthrough information. In *Information Retrieval Technology* (pp. 593-600).Springer Berlin Heidelberg.

Lee, U., Liu, Z., & Cho, J., 2005.Automatic identification of user goals in web search.In*Proceedings of the 14th international conference on World Wide Web* (pp. 391-400). ACM.

Lewandowski, D., 2006. Query types and search topics of German Web search engine users. *Information Services and Use*, 26(4), 261-269.

Lewandowski, D., Drechsler, J., & Mach, S., 2012. Deriving query intents from web search engine queries. *Journal of the American Society for Information Science and Technology*, 63(9), 1773-1788.

Mendoza, M., & Zamora, J., 2009. Identifying the intent of a user query using support vector machines. In *String Processing and Information Retrieval* (pp. 131-142). Springer Berlin Heidelberg.

Morrison, J. B., Pirolli, P., & Card, S. K., 2001, March. A taxonomic analysis of what World Wide Web activities significantly impact people's decisions and actions. In *CHI'01 extended abstracts on Human factors in computing systems* (pp. 163-164). ACM.

Rose, D. E., Levinson, D., 2004. Understanding user goals in web search. In *Proceedings of the 13th international conference on World Wide Web* (pp. 13-19). ACM.

Wu, D., Zhang, Y., Zhao, S., & Liu, T., 2010. Identification of Web Query Intent Based on Query Text and Web Knowledge. In Pervasive Computing Signal Processing and Applications (PCSPA), 2010 First International Conference on (pp. 128-131). IEEE.