International Journal of Computer Applications (0975 - 8887)
Volume 75 - No. 17, August 2013

Search Engines going beyond Keyword Search: A Survey

Mahmudur Rahman
School of Computing and Information Sciences
Florida International University,
Miami, FL 33199

ABSTRACT

In order to solve the problem of information overkill on the web or large domains, current information retrieval tools especially search engines need to be improved. Much more intelligence should be embedded to search tools to manage the search and filtering processes effectively and present relevant information. As the web swells with more and more data, the predominant way of sifting through all of that data —keyword search —will one day break down in its ability to deliver the exact information people want at our fingertips. Hence search engines are trying to break the shackles of the concept of keyword search what typically most search engines do. This paper tries to identify the major challenges for today's keyword search engines to adapt with the fast growth of web and support comprehensive user demands in quick time. Then it surveys different non-keyword based paradigms proposed, developed or implemented by researchers and different search engines and also classifies those approaches according to the features focused by the different search engines to deliver results.

General Terms:

Search Engine, Information Retrieval

Keywords:

keyword based search, semantic web search, search engine, computational knowledge engine, question-answering system

1. INTRODUCTION

Search is one of the keys to the web's success. Search engines have forever changed the way people access and discover knowledge, allowing information about almost any subject to be quickly and easily retrieved within seconds. Indeed, the main way people access the web is via that wee box that seems to read their mind from a few words and return a list of links to resources they want. This approach has become so successful to finding information that on the one hand it is difficult to remember how people managed to find any information at all prior to web based keyword search, and on the other hand, it is difficult to envision needing or wanting any other tool for information discovery. Successful paradigms can sometimes constrain one's ability to imagine other ways to ask questions that may open up new and more powerful possibilities. It enables one to do so much information discovery that it is difficult to imagine what she cannot do with the paradigm of continually refining search terms to get help a busy person find a better job quickly, effectively, that is a match for her passion and skills. And if that person could use some extra training to support that skill to get that better job, how would the Google paradigm bring in that highly relevant information that is outside the constraints of the keyword search?

In the Information Retrieval and Information Seeking literature, these kinds of more complex, rich information discovery and knowledge building tasks have been modeled in terms of search strategies and tactics. Today's web consists of various types of data and the search engines need to provide the data according to the user's query. The demand for providing exploration facility to users rather than keyword search is getting stronger day by day and search engines are trying to add new dimension or features to support these alternate kinds of search and knowledge building. Also the search engine tends to capture the semantic web trend which seeks to improve search accuracy by understanding searcher intent and the contextual meaning of terms as they appear in the searchable data space, whether on the web or within a closed system, to generate more relevant results

Our work investigates schemes and approaches, outside the boundary of text or keyword search, developed and implemented by different search engines and researchers. The main challenges for our work is that the number of different approaches, being developed and implemented, is so vast that it is very hard to analyze and classify all those paradigms. The challenges for non-keyword based search are three-fold. First, it is still the early stage for the search engines to go for the semantic search or exploratory search. Second, it is not obvious that the search performance of non-keyword based search outperforms the keyword based search while doing empirical evaluation. Third, although exploratory search tries to go beyond keyword search offering visualization, user intent capture, visual query, they still depend on keyword based search.

This paper is organized as follows. Section 2 defines keyword search and lists some of its problems and Section 3 details related works. Section 4 provides an overview of different types of search strategies while Section 5 discusses about the technical challenges faced by the keyword based search engines. Section 6 classifies different approaches of search engines, Section 7 discusses some ideas and Section 8 concludes.

2. PROBLEM DEFINITION

2.1 Keyword Searching

This is the most common form of text search on the web. Most search engines do their text query and retrieval using keywords. What is a keyword, exactly? It can simply be any word on a web page. Basically, a keyword is an index entry that identifies a specific record or document. Unless the author of the web document specifies the keywords for her document, it is up to the search engine to determine them. Essentially, this means that search engines pull out and index words that appear to be significant.

Since search engines are software programs, not rational human beings, they work according to rules established by their creators for what words are usually important in a broad range of documents. The title of a page, for example, usually gives useful information about the subject of the. Words that are mentioned towards the beginning of a document are given more weight by most search engines. The same goes for words that are repeated several times throughout the document. Some search engines index every word on every page. Others index only part of the document.

2.2 The Problem with Keyword Searching

Keyword searches have a tough time distinguishing between words that are spelled the same way, but mean something different (i.e. hard cider, a hard stone, a hard exam, and the hard drive on your computer). This often results in hits that are completely irrelevant to our query. Some search engines also have trouble with so-called stemming i.e., if you enter the word "big", should they return a hit on the word, "bigger"? What about singular and plural words? What about verb tenses that differ from the word you entered by only an "s", or an "ed"? Search engines also cannot return hits on keywords that mean the same, but are not actually entered in your query. A query on heart disease should not return a document that used the word "cardiac" instead of "heart". So search engines need to understand searcher intent and the contextual meaning of terms as they appear in the searchable data space, whether on the web or within a closed system, to generate more relevant results. In short, existing keyword based search engines need to be more intelligent and comprehensive.

3. RELATED WORKS

Information seeking, retrieval, discovery and analysis, especially search engine paradigms and performance evaluation, have been a very active area of research. As the web is growing with more and more data, it is hard for current keyword based search engines to deliver the exact information at our fingertips in the way of sifting all of the enormous volume of data. Also as people are getting more involved with multimedia data like audio, video, image and number of information available for those type of information is increasing day by day, search engines are going to support content based search rather than just text-based search. In fact, some argue that keyword search is already delivering diminishing returns. So different approaches and strategies are being developed and implemented by both researchers and search engine companies.

For instance, several content-based image retrieval systems allow a user to sketch a coarsely detailed picture and retrieve similar images based on color, texture, and shape similarities (e.g. [25], [27], [4]). Chang et al. [5] propose a novel, interactive system on the web, based on the visual paradigm, with spatiotemporal attributes. It is the first on-line video search engine supporting automatic object based indexing and spatiotemporal queries. Another significant work was done by IBM researchers [12] when they proposed the QBIC system, which relies on query on image and video content. When Yahoo.com and Google.com image search support queries by keyword, size, coloration, file type, and domain; QBIC search engine provides especial query methods like Example images, Sketches and drawings, User-selected color and texture patterns, Camera and object motion. Russian museum's online digital collection uses QBIC engine. Funkhouser et al. [13] propose a web-based search engine system that supports queries based on 3D sketches, 2D sketches, 3D models, and/or text keywords. For the shape-based queries, a new matching algorithm has been developed that uses spherical harmonics to compute discriminating similarity measures without requiring repair of model degeneracy or alignment of orientations.

Traditional keyword/text based search lacks understanding of the user's intent and the web's content both queries and documents are typically treated as a word, missing semantic-level understanding. The solution of improving search accuracy is by understanding searcher intent and the contextual meaning of terms. Dittenbach et al. [10] presents ConceptWorld, an instrument to automatically discover various facets of a topic of interest by extracting concepts from Web documents. The result materializes as a network of semantic concepts with their various contextual interrelations and provides a holistic view on the topic of interest. Liu et al. [19] introduced semantic web technologies to e-commerce search field, and designed a semantic network structure for the new search system, and discussed the key technologies in e-commerce semantic search, such as semantic structure and semantic search algorithm. Compared with traditional search, semantic search can return more relevant semantic information and can extract users' search input more accurate. Zou et al. [36] propose and implement a semantic search prototype system. The experimental results show that semantic expansion search by proposed methodology can overcome limitations in comparison with traditional keyword search mode, and achieve higher recall ratio and precision ratio. Lee and Tsai [18] design an interactive semantic search engine which collects feedback by means of selection in order to better capture users personal concepts. Chiang et al. [7] present a semantic search engine based on the smart web query (SWQ) method for web data retrieval. The SWQ architecture contains three main parts: SWQ search engine and its subcomponents: "query parser" and "context ontology determination engine"; context ontologies for domains of application; a semantic search filter which is to improve search precision based on retrieving term properties in context ontologies. Bhagwat and Polyzotis [3] propose a semantic-based file system search engine —Eureka, which uses an inference model to build the links between files and a FileRank metric to rank the files according to their semantic importance. Kandogan et al. [17] develop a semantic search engine Avatar, which combines the traditional text search engine with use of ontology annotations [17]. Avatar has two main functions: (1) extraction and representation (2) interpretation a process of automatically transforming a keyword search to several precise searches.

Swoogle [9] is a crawler-based semantic search engine. Three main functions are provided by Swoogle, which are finding appropriate ontologies for specific terms involved; finding instance data semantic web documents defined by specific classes and properties; characterizing the semantic web by gaining interrelationship among metadata in semantic web, Swoogle is able to answer the questions about semantic web structure. Bhagdev et al. [2] describe hybrid search, a search method supporting both document and knowledge retrieval via the flexible combination of ontology based search and keyword-based matching. Hybrid search smoothly copes with lack of semantic coverage of document content, which is one of the main limitations of current semantic search methods. Also it is shown how the method outperforms both keyword-based search and pure semantic search in terms of precision and recall in a set of experiments performed on a collection of about 18,000 technical documents. Wilson et al. [33] aims to exploratory search by focusing on the techniques and visualizations that allow users to interact with and have control over their findings. They have shown that there is substantial room for improving the support provided to users who are exhibiting more exploratory forms of search, including when users may need to learn, discover, and understand novel or complex topics.

4. SEARCH CLASSIFICATION

Information retrieval based on usual search scenario involves someone typing in a query to a search engine and receiving answers in the form of a list of documents in ranked order. Although searching the World Wide Web (web search) is by far the most common application involving information retrieval, search is also a vital part of applications in corporations, government and many other domains. **Vertical search** is a specialized form of web search where the domain of the search is restricted to a particular topic. **Enterprise search** is the practice of making content from multiple enterprise-type sources, such as databases and intranets, searchable to a defined audience. It indexes data and documents from a variety of sources such as: file systems, intranets, document management systems, e-mail, and databases. **Desktop search** is the personal version of enterprise search, which search the contents of a user's own computer files, rather than searching the Internet. These tools are designed to find information on the user's PC, including web browser histories, e-mail archives, text documents, sound files, images and video. Desktop search is emerging as a concern for large firms for two main reasons: untapped productivity and security. **Peer-to-peer search** involves finding information in networks of nodes or computers without any centralized control. This type of search began as a file sharing tool for music but can be used in any community based on shared interests or even shared locality in the case of mobile devices.
The term "search engine" is often used generically to describe crawler-based search engines, human-powered directories, and hybrid search engines. These types of search engines gather their listings in different ways, through crawler-based searches, human-powered directories, and hybrid searches.

4.1 Crawler-based search engines

Crawler-based search engines, such as Google, create their listings automatically. They "crawl" or "spider" the web, then people search through what they have found. If web pages are changed, crawler-based search engines eventually find these changes, and that can affect how those pages are listed. Page titles, body copy and other elements all play a role.

4.2 Human-powered directories

A human-powered directory, such as the Open Directory Project [24] depends on humans for its listings. (Yahoo!, which used to be a directory, now gets its information from the use of crawlers.) A directory gets its information from submissions, which include a short description to the directory for the entire site, or from editors who write one for sites they review. A search looks for matches only in the descriptions submitted. Changing web pages, therefore, has no effect on how they are listed. Techniques that are useful for improving a listing with a search engine have nothing to do with improving a listing in a directory. The only exception is that a good site, with good content, might be more likely to get reviewed for free than a poor site. Yahoo Directory is another best known directory site.

4.3 Hybrid search engines

Today, it is extremely common for crawler-type and human-powered results to be combined when conducting a search. Usually, a hybrid search engine will favor one type of listings over another. For example, MSN Search is more likely to present human-powered listings from LookSmart [21]. However, it also presents crawler-based results, especially for more obscure queries. Other good examples are DogPile [26] and MetaCrawler [31].

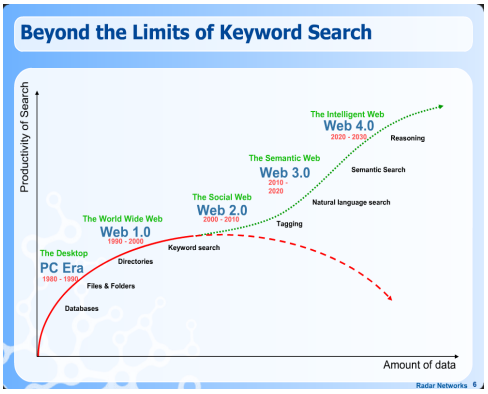


Fig. 1. Search engine going beyond keyword search.

5. CHALLENGES FOR TODAY'S SEARCH ENGINE

One of the greatest human needs that have evolved in the 21st century is the need to know as much as possible about something before making a decision. Today, the fear that there's knowledge out there that could be relevant to the decisions, and that people are not using it and getting a lesser deal, haunts us all. There are facts, figures, opinions, comments, user reviews. Information, unlike wealth, has grown directly in proportion to its usage. This information fire hose impacts both individuals and enterprises. While individuals crave to know as much as possible before committing to something, enterprises find their customers more demanding or their competition more informed. Staying on top of this complex, voluminous information tidal wave has become crucial for survival. As a response to this, search companies have sprung up, with Google in the lead. For about a decade now, search engines of all sorts are battling it out with terabytes of content on the Internet. They are facing various challenges as follows: **Firstly**, the Internet (and other networks within or without the enterprise) is moving from being information stores to knowledge networks. As the volume and complexity of knowledge grows, search is becoming inadequate. Search companies are losing ground fast. **Secondly**, search engines need to consider the issue findability carefully which is becoming crucial day by day. Search as a tool is fine for information stores, but poor for knowledge bases. Knowledge bases need to have findability. While, a lot of people confuse findability with search, the two are really not the same thing. Search tries to solve the problem of locating information that people already know exists somewhere in a corpus. Findability encompasses search, but also deals with the problem of how to make the searcher aware of other relevant information, that they didn't know existed in the corpus. No semantic or index based search can ever completely fill this gap. A good approach to solving this problem would be to marry a social-tagging system such as de.li.ci.ous or digg and a semantic analysis engine. The Findability solution would need to work as a facilitator that allows people to share their personal experiences and knowledge around a product and build a knowledge community.
Thirdly, despite the visual nature of the web, few search engines have focused on retrieving visual information such as images and videos. Indeed, while there has been some success in developing search engines for text, search engines for other media on the Web (images, audio, and video) are still rare and not as powerful. Visual information is published both embedded in Web documents and as stand-alone objects. It takes the form of images, graphics, bitmaps, animations, and videos. There is thus a vital need for Web multimedia search engines focusing on content rather than purely keyword or text. Such engines are useful for many applications, including law enforcement; image

copyright protection, filtering of inappropriate mature content, criminal tracking, home entertainment, education, and training. **Lastly**, as the Web swells with more and more data day by day, the predominant way of sifting through all of that data keyword search will one day break down in its ability to deliver the exact information people want at our fingertips. In fact, some argue that keyword search is already delivering diminishing returns as the figure 1 above by Nova Spivack [28] implies. There are many approaches being tried: social search, tagging, guided search, natural-language search, statistical methods, open search, semantic search, and (way out there) artificial intelligence. They all have their problems. Tags are too messy and inconsistent. Natural-language requires too much computing power, is difficult to scale, and does not deal with structured data well. Semantic search is perhaps the most promising, but it essentially requires every single Webpage to be re-written. But this search is in early stage and there are plenty of scopes to develop.

6. SEARCH ENGINES GOING BEYOND KEYWORD SEARCH

The number of different approaches, being developed and implemented, is vast enough to include all of them in a single paper. Still we have tried to classify the existing significant search engine approaches as the following categories.

6.1 Content based search

Searching the web for specific information has become a very time consuming and inefficient task for even the most expert users. Image is worth a thousand words, and object for object, pictures are several orders of magnitude larger and more subtle information carriers than written language. Video, a dense stream of still images, increases the difficulty of information retrieval several magnitudes beyond text or even individual images or graphics. But content-based image retrieval is critical to many applications, and similar in principle to much text retrieval, and useful image search systems are emerging from such vendors as IBM, Excalibur, and Virage. IBM's image management system, Query by Image Content (QBIC) [12], provides searching of still graphics and video collections based on properties such as shape, texture, sketches, and other attributes. IBM supplements its pure image searching with text searching. Instead of inverted indices pointing to occurrences of words, Excaliburs [16] neural network technology uses what it calls Adaptive Pattern Recognition Processing (APRP). Excalibur's recent acquisition of Interpix Software Corporation and deals with Yahoo has extended its multimedia search and retrieval reach to Web servers. VIRAGE's Image Search engine Library [1] is a static or dynamically linkable library which offers primitive functions that can be used to enter images into a searchable collection and then to query that collection. What distinguishes the Virage engine is its efficiency and precision in managing image attributes. The four primary attributes in a Visage image collection are color distribution, color placement, structure, and texture.
Another good example of content based search engine is TinEye Reverse Image Search [27]. TinEye is a reverse image search engine. One can submit an image to TinEye to find out where it came from, how it is being used, if modified versions of the image exist, or to find higher resolution versions. TinEye regularly crawls the web for new images, and also accepts contributions of complete online image collections. Piximilar Visual Search [4] is an API that allows us to search through large image collections, without keywords or metadata, to instantly retrieve visually similar images. It can be used in combination with keywords to refine searches on extremely large collections. Another content based search engine is XIRS [29] (shown in figure 2(a)) which is an XML-based Image Retrieval System where a user request can be an image file or a keyword. The CBIR (Content Based Image Retrieval) system and the current search engines (e.g. Google, Yahoo.) make image search possible only when the query is a keyword. This type of search is limited because keywords are not expressive enough to describe all important characteristics of an image. For example, an exact match request cannot be formulated in such systems. Thus, a search system is proposed in which a request might be an image file or a keyword. The MPEG-7 standard is used for describing an image as an XML document. Content-based visual queries have been primarily focused on still image retrieval. VideoQ [5] is the first on-line video search engine supporting automatic object based indexing and spatiotemporal queries.It supports object-oriented content-based video search along with keyword-based search. Indexing video objects with motion attributes and developing good spatiotemporal metrics have been the key issues in this paradigm. The system performs well, with the user being able to retrieve complex video clips such as those of skiers and baseball players with ease.

6.2 Question-Answering (QA) system

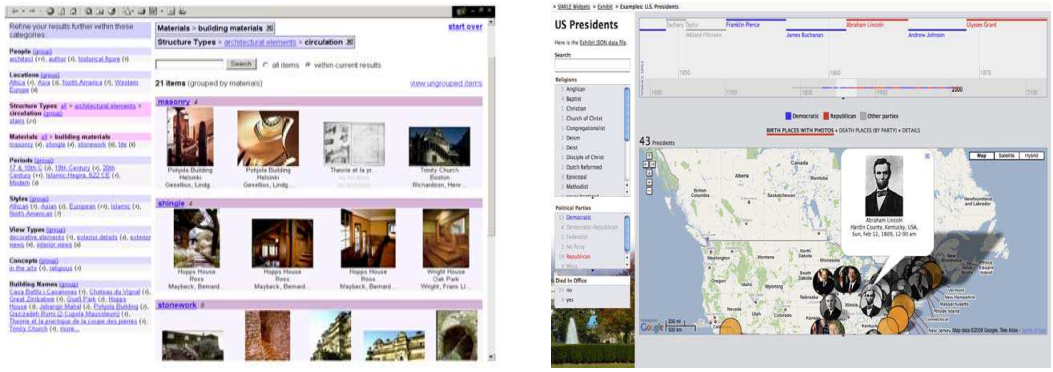
The goal of the QA process is to retrieve answers to questions rather than full documents or best-matching passages, as most information retrieval systems currently do. Most question answering systems use a combination of techniques from computational linguistics, information retrieval and knowledge representation for finding answers.
A search engine is a system which partially process question answering. Search engines do not have an important capability-deduction capability, the capability to synthesize an answer to a query by drawing on bodies of information which reside in various parts of the knowledge base. In these days, the search engine is trying to move to serve question answering going beyond its typical keyword based search. "Yahoo! Answers" is such a big example. It is a place where people ask and answer questions on any topic. In addition, Yahoo! Answers facilitates the preservation and retrieval of answered questions aimed at building an online knowledge base. The site is meant to encompass any topic, from travel and dining out to science and mathematics to parenting. Despite its relative novelty, Yahoo! Answers already features more than 10 million questions and a community of several million users as of February 2007. In addition, other recently launched services, like Microsoft's Live QnA and Amazon's Askville, seem to follow the same basic interaction model. Another approach to Question-answering system is provide system generated direct answers to the question rather than giving answers from the members of the system. One good example is Wolfram—Alpha [11]. Suppose the question is "Who is Barack Obama?" And the Wolfram system directly gives the answer-politician, then basic information including place and date of birth, leadership position and other valuable information at once rather than providing a list of documents containing the keyword "Barach Obama" in keyword based search engine. This is indeed a significant improvement of search engine over traditional keyword based search.

6.3 Computational knowledge engine

This is one of the powerful implementations of semantic search applications and the biggest example is the successful Wolfram—Alpha [11] (shown in figure 2(b)). It introduces a fundamentally new way to get knowledge and answers not by searching the web, but by doing dynamic computations based on a vast collection of built-in data, algorithms, and methods. Its goal is to accept completely free-form input, and to serve as a knowledge engine that generates powerful results and presents them with maximum clarity. Wolfram Mathematica 8 pioneers free-form linguistic input, allowing users to enter plain English and get immediate results and the Mathematica input for further exploration without the need for syntax. It is a breakthrough in

(a) (b)

Fig. 2. (a) XIRS interface. (b) Wolfram—Alpha Mathematica combines Knowledge and Computation.



(a) (b)

Fig. 3. (a) The Flamenco interface permits users to navigate by selecting from multiple facets. In this example, the matching images are grouped by subcategories of the Materials facet's selected Building Materials category. (b) The Exhibit faceted search interface takes a slightly different approach, only filtering facets without a selection, so that previous selections can be seen in the context of the options at the time.

usability that makes many programming and development tasks as easy as entering a query in English.

6.4 Semantic web search

A semantics search engine attempts to make sense of search results based on context. It automatically identifies the concepts structuring the texts. For instance, if one search for "election" a semantic search engine might retrieve documents containing the words "vote", "campaigning" and "ballot", even if the word "election" is not found in the source document. An important part of this process is disambiguation, both of the queries and of the content on the web. What this means is that the search engine through natural language processing will know whether you are looking for a car or a big cat when you search for "jaguar".

Semantic search has the power to enhance traditional web search, but it will not replace it. A large portion of queries are navigational and semantic search is not a replacement for these. Research queries, on the other hand, will benefit from semantic search. The four search engines below all use semantic analysis to sift through and present data. But, as anyone will see, they do not do this in the same way The first one is Hakia [14], which is a general purpose semantic search engine, as opposed to e.g. Powerset and SenseBot (described below), that search structured corpora (text) like Wikipedia. For some queries (typically popular queries and queries where there is little ambiguity), Hakia produces resumes. These are portals to all kinds of information on the subject. Every resume has an index of links to the information presented on the page for quick reference. The elements of these resumes will vary according to the nature of the query (e.g. biography, bibliography, timeline etc. for persons, government, economy, culture etc. for countries). Resumes are excellent for researching a topic.

The second impressive semantic engine is SenseBot [30]. It is a web search engine that summarizes search results into one concise digest on the topic of the query. The search engine attempts to understand what the result pages are about. For this purpose it uses text mining to analyze web pages and identify their key semantic concepts. The summary serves as a digest on the topic of the query, blending together the most significant and relevant aspects of the search results. It contains a tag cloud, relating your query to other relevant concepts and a list of sentences believed to define or describe your query. Each sentence is followed by a link to the source. Not all of the summaries are informative or even intelligible, but that is likely to improve; Like Hakia, SenseBot is in beta.

Powerset [22] is another semantic search engine focused on natural language processing. In other words, Powerset will not search based simply on keywords alone, but will try to understand the semantic meaning behind the search phrase as a whole. The company launched in May 2008 with intentions of making search more easy and intuitive. Microsoft buys Powerset, gets foot in semantic search door. Powerset is at present not a regular web search engine. It works best on smaller, relatively structured corpora. The technology offers a comprehensive view of such information. On the search results page, Powerset often answers questions directly. The vital feature is the way it aggregates information from across multiple articles.

DeepDyve [8] is a powerful, professional research tool available for free for the general public. It is a research engine that lets

us access expert content from the "Deep Web", the part of the Internet that is not indexed by traditional search engines (e.g. databases, journals etc.). The search results are presented in a complex manner with many advanced options for refining, sorting or saving the search. Despite the complexity, the search results are relatively easy to navigate.

6.5 Exploratory Search

In standard web search, users submit a query via a search box and view a textual list of results. More recently, a new class of search has emerged, called exploratory search [6], which supports the exploration and discovery of information through both querying and browsing strategies. In that regard, Marchionini [20] identified three types of search activities: (1) lookup, (2) learn and (3) investigate. Lookup searches can be thought of as traditional search, while learn and investigate searches relate to discovery-oriented tasks. In recent years, a few desktop exploratory search systems have been proposed in the literature. Yee et al. [34] presented an alternative interface for exploring large collections of images using hierarchical faceted metadata and dynamically generated query previews. Tvaroek & Bielikov [32] proposed a personalized faceted browser that facilitates exploratory search by providing users with an integrated search and navigation interface that combines full text, faceted, content-based and collaborative search. Now we will describe some of the exemplary exploratory search strategies as follows.

6.5.0.1 Faceted search. Faceted search, also called faceted navigation or faceted browsing, is a technique for accessing a collection of information represented using a faceted classification, allowing users to explore by filtering available information. Each facet typically corresponds to the possible values of a property common to a set of digital objects. This approach permits existing web-pages, product descriptions or articles to have this extra metadata extracted and presented as a navigation facet.
Flamenco [34] (shown in figure 3(a)) is a clear example of the features provided by faceted search using multiple hierarchical facets. Providing interfaces to fixed collections, including art, architecture, and tobacco documents, Flamenco presents faceted hierarchies to produce menus of choices for navigational searching. With Flamenco, users were more successful at finding relevant images (for the structured tasks) and reported higher subjective measures (for both the structured and exploratory tasks).
Huynh et al. [15] has developed Exhibit(shown in figure 3(b)), a faceted search system that is similar to flamenco in many ways, but has some significant developments. One key advance is that, where a selection in Flamenco filters all the facets, a selection in Exhibit filters all the other facets and leaves the facet with the selection unchanged. This provides two benefits: first, users can easily change their selection and second, users can make multiple selections in one facet. This allows users to see, for example, the union of all the red and blue clothes, rather than just red or just blue. This support for multiple selections within a single facet has been recently added to ebay.com, but remains unavailable in services such as Google Product Search.
The Relation Browser [35] (shown in Figure 4(a)) takes another approach to the faceted search. One notable difference is that multiple selections lead to their intersection of results being displayed. Another feature that the Relation Browser provides is a preview of the effect of clicking has on other facets. Graphical representations behind each item in each facet show how many documents can be found by selecting it. This technique revives the query preview strategy, which is a helpful alternative to the simple numeric volume indicators that are included in most classification-based systems.

6.5.0.2 Community-driven classification system. There is a growing consensus that tag clouds used in social classification are more valuable for users who are making sense of information, rather than for finding specific information. Some systems are trying to take the benefits of tagging to produce what are known as folksonomies or community-driven classification systems. One approach, used by the MrTaggy [23] interface (shown in Figure 4(b)) allows users to perform searches based entirely on community-generated tags by including or excluding tags from a list of related tags . Searches are initiated with a pair of selections from two tag clouds: one containing adjectives and the other containing nouns and other objects.

7. DISCUSSIONS

In previous sections, we have discussed many non-keyword based search approaches but is it possible to predict the future search engine trend? What is about the content space in future in web? We will try to give some ideas to these questions.

7.1 The Search Engine of the future

We think that the search engine of the future will be to some extent semantic. May be it will combine some of the features of exploratory search and content based search. The way we see things, the semantic search engines will harvest content words in much the same way as is done today. They will also weight in much the same way as today, but when today's search engine stops and present the results, the semantic search engine will go on some steps further. The semantic search engine will analyze the collection of content words, the relative weight, the cohesion between them and the way they are (semantically) connected. The search engine will then find other pages or collections of pages with the same semantic profile or with a semantic profile that falls within an acceptable threshold of values.

7.2 Importance of Content

Content will be even more important tomorrow than it is today, and the way people write contents will be essential. Today it is important to know the relevant and realistic keywords and then optimize pages to hit high for these words. Tomorrow the content will need to be much more varied where keywords are not enough. Also synonyms, acronyms, alternatives, opposites and variations are necessary. The scope of the content in the entire site will increase in importance. The content space of the site will therefore increase in importance. They need to support content based search approach for efficient and comprehensive retrieval of multimedia contents and representations.

8. CONCLUSIONS

In this paper, we have presented different search paradigms of various search engines like semantic search or concept based search, exploratory search, content based search, open domain question-answering and specific trend of semantic search like computation knowledge engine. Also various issues, drawbacks and challenges of today's keyword based search engines have been discussed. Semantic search seems to be a promising technology but might have set the expectations way too high. Although some systems have appeared in recent years, more work needs to be done in this area and a number of questions remain to be addressed. First, more studies must be carried out on users and their queries in order to understand them in a more full manner and to meet their needs. Second, more efforts are needed in the area of narrowing the semantic gap in order to allow people to retrieve exact relevant information emphasizing concept based search. Third, there is a vital need for standardization in each of multimedia (audio,image,video) description, which would allow search engines to retrieve content with more precision. Fourth, comprehensive and efficient indexing and retrieval techniques should be developed to deal specifically with the great number of high-dimensional features that needs to be handled in the Web

(a) (b)

Fig. 4. (a) The Relation Browser interface provides consistent facets, where the list of values is not filtered by selections. Instead, users can see the reduction in files associated with each facet-value with the bar-chart style visualizations. (b) The MrTaggy interface allows users to include or exclude tags from their search instead of using keywords.

context. Other important issues that remain to be addressed include ensuring a better coverage of the Web, and integration of different search paradigms.

9. REFERENCES

[1] Jeffrey R. Bach, Charles Fuller, Amarnath Gupta, Arun Hampapur, Bradley Horowitz, Rich Humphrey, Ramesh C. Jain, and Chiao-Fe Shu. Virage image search engine: an open framework for image management. pages 76–87, 1996.

[2] Ravish Bhagdev, Sam Chapman, Fabio Ciravegna, Vitaveska Lanfranchi, and Daniela Petrelli. Hybrid search: effectively combining keywords and semantic searches. In *Proceedings of the 5th European semantic web conference on The semantic web: research and applications*, ESWC'08, pages 554–568, Berlin, Heidelberg, 2008. Springer-Verlag.

[3] Deepavali Bhagwat and Neoklis Polyzotis. Searching a file system using inferred semantic links. In *Proceedings of the sixteenth ACM conference on Hypertext and hypermedia*, HYPERTEXT '05, pages 85–87, NY, USA, 2005. ACM.

[4] Piximilar: Image by color. `http://research.cs.wisc.edu/vision/piximilar/`.

[5] Shih-Fu Chang, W. Chen, H.J. Meng, H. Sundaram, and Di Zhong. A fully automated content-based video search engine supporting spatiotemporal queries. *Circuits and Systems for Video Technology, IEEE Transactions on*, 8(5):602–615, 1998.

[6] Shih-Fu Chang, Lyndon S. Kennedy, and Eric Zavesky. Columbia university's semantic video search engine. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, CIVR '07, pages 643–643, NY, USA, 2007. ACM.

[7] Roger H. L. Chiang, Cecil Eng Huang Chua, and Veda C. Storey. A smart web query method for semantic retrieval of web data. *Data Knowl. Eng.*, 38(1):63–84, July 2001.

[8] DeepDyve. `http://www.deepdyve.com/`.

[9] Li Ding, Tim Finin, Anupam Joshi, Rong Pan, R. Scott Cost, Yun Peng, Pavan Reddivari, Vishal Doshi, and Joel Sachs. Swoogle: a search and metadata engine for the semantic web. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, CIKM '04, pages 652–659, NY, USA, 2004. ACM.

[10] M. Dittenbach, H. Berger, and D. Merkl. Automated concept discovery from web resources. In *Web Intelligence, 2006. WI 2006. IEEE/WIC/ACM International Conference on*, pages 309–312, 2006.

[11] WolframAlpha: Computational Knowledge Engine. `http://www.wolframalpha.com/`.

[12] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Qian Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker. Query by image and video content: the qbic system. *Computer*, 28(9):23–32, 1995.

[13] Thomas Funkhouser, Patrick Min, Michael Kazhdan, Joyce Chen, Alex Halderman, David Dobkin, and David Jacobs. A search engine for 3d models. *ACM Trans. Graph.*, 22(1):83–105, January 2003.

[14] Hakia. `http://company.hakia.com/`.

[15] David F. Huynh, David R. Karger, and Robert C. Miller. Exhibit: lightweight structured data publishing. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 737–746, NY, USA, 2007. ACM.

[16] IBM. Informix excalibur text search datablade. `http://www-01.ibm.com/software/data/informix/blades/excaliburtext/`.

[17] Eser Kandogan, Rajasekar Krishnamurthy, Sriram Raghavan, Shivakumar Vaithyanathan, and Huaiyu Zhu. Avatar semantic search: a database approach to information retrieval. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, SIGMOD '06, pages 790–792, NY, USA, 2006. ACM.

[18] Wei-Po Lee and Tsung-Che Tsai. An interactive agent-based system for concept-based web search. *Expert Systems with Applications*, 24(4):365 – 373, 2003.

[19] Zhusong Liu and Yuqin Zhang. Research and design of e-commerce semantic search. In *Information Management, Innovation Management and Industrial Engineering (ICIII), 2010 International Conference on*, volume 4, pages 332–334, 2010.

[20] Gary Marchionini. Exploratory search: from finding to understanding. *Commun. ACM*, 49(4):41–46, April 2006.

[21] LookSmart Search Marketing. `http://search.looksmart.com/`.

[22] Paul Miller. Powerset shows semantic search solution. http://www.zdnet.com/blog/semantic-web/powerset-shows-semantic-search-solution/141, May 2008.

[23] MrTaggy. http://mrtaggy.com/.

[24] Open Directory Project. http://www.dmoz.org/about.html.

[25] RevIMG. http://www.revimg.com/.

[26] Dogpile Web Search. http://www.dogpile.com/.

[27] TinEye Reverse Image Search. http://www.tineye.com/.

[28] Nova spivack. The road to semantic search the twine.com story. http://www.novaspivack.com/uncategorized/the-road-to-semantic-search-the-twine-com-story, December 2009.

[29] G. N. Fanzou Tchuissang, Xu De, N. Wang, and François Siewe. Xirs: an xml-based image retrieval system. In *Proceedings of the 7th Conference on 7th WSEAS International Conference on Multimedia, Internet & Video Technologies - Volume 7*, MIV'07, pages 233–238, Stevens Point, Wisconsin, USA, 2007. World Scientific and Engineering Academy and Society (WSEAS).

[30] SenseBot: The Search Engine that finds sense in a heap of Web pages. http://www.sensebot.net/.

[31] MetaCrawler: Search the Search Engines. http://www.metacrawler.com/.

[32] Michal Tvarožek and Mária Bieliková. Collaborative multi-paradigm exploratory search. In *Proceedings of the hypertext 2008 workshop on Collaboration and collective intelligence*, WebScience '08, pages 29–33, NY, USA, 2008. ACM.

[33] Max L. Wilson, Bill Kules, m. c. schraefel, and Ben Shneiderman. From keyword search to exploration: Designing future search interfaces for the web. *Found. Trends Web Sci.*, 2(1):1–97, January 2010.

[34] Ka-Ping Yee, Kirsten Swearingen, Kevin Li, and Marti Hearst. Faceted metadata for image search and browsing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '03, pages 401–408, NY, USA, 2003. ACM.

[35] Junliang Zhang and Gary Marchionini. Evaluation and evolution of a browse and search interface: Relation browser++. In *Proceedings of the 2005 national conference on Digital government research*, dg.o '05, pages 179–188. Digital Government Society of North America, 2005.

[36] Guobing Zou, Bofeng Zhang, Yanglan Gan, and Jianwen Zhang. An ontology-based methodology for semantic expansion search. In *Fuzzy Systems and Knowledge Discovery, 2008. FSKD '08. Fifth International Conference on*, volume 5, pages 453–457, 2008.