Computing, Information Systems & Development Informatics Journal
Vol. 8 No. 4, December, 2017 - www.cisdijournal.org

A Systematic review of Information Retrieval and Ranking Algorithms

Yaya, J.O & Robert, A.B.C.
Department of Computer Science
University of Ibadan
Ibadan, Nigeria

ABSTRACT

The information age is characterized by the huge amount of data available to people. The exact amount of data in the globe was estimated at about 4.4 zettabytes as at 2013 and this is assumed to rise to about 44 zettabytes by 2020. Arguably, a zettabyte is equivalent to 44 trillion gigabytes. The Internet is regarded as the largest source of data, an enormous source of information with a great stock of various web pages & hyperlinks. The information on the internet can be queried with the aim of enabling users to find whatever they want. In the same vein, the number and diversity of internet users continue to increase. Millions of users all over the world search the Web on a daily basis with the aim of finding the right answer or solution to problems. The Web has revolutionized access to digitally available data; as such, web information search and retrieval have become key aspects of human interactions. Information retrieval is a complex process, the study of which has attracted a lot of research efforts. This paper carried out a systematic review of information retrieval, search, andranking algorithms. The objective is to provide a theoreicalk framework and identify research gaps that can serve as a basis for the development of an enhanced recursive model for individualized search

Keywords: Information, Retrieval, Recursive Model, Individualized Search & Ranking

CISDI Journal Reference Format
Yaya, O.J & Robert,. A.B.C. (2017): A Systematic review of Information Retrieval and Ranking Algorithms. Computing, Information Systems & Development Informatics Journal. Vol. 8 No. 4. Pp 58-95
Available online at www.cisdijournal.net

1. INTRODUCTION

1.1 Information Retrieval

Information retrieval (**IR**): This is seen as an act of retrieving information specific to user need from a pool of existing information resources. As such, it is seen as a systemic search of body of information within the documents themselves, and search metadata that describes data. An information retrieval process begins with the entry of a query by a user into the system. Queries is formal statements of information search, for instance, searching strings within the web search engine. It is important to note that queries does not necessary identifies objects in a data collection process. Here the object can be an entity that can be presented in the form of content information to which user queries are matched. The data objects can be in the form of image, textual documents depending on its application (Goodrum, 2000), mind maps (Beel et al., 2009) audio (Foote, 1999), or even videos. The process involves matching the user queries with database information. This is in contrast with the classical SQL queries where the results returned does not always match the query. IR system returns several objects matching the query, with varying degrees of relevancy. So the results are usually ranked based on some measure of relevancy to the user query.

The numeric score computes how well the objects matches the various queries, through ranking each of the object based on their score, as such, the best ranked object are reveled to the users, where the processes can often be iterated when the users feels redefine the query (Frakes, 1992). As argued by Jansen & Rieh (2010), the major difference IR searching and database searching lies in the ranking of results.

In the current era characterized by the sheer volume of available information, automated IR system is used in reducing the so-called "information overload" (Jansen & Rieh, 2010). Interestingly, searches in the IR systems could be content based or full text indexing. Examples of IR systems include online databases, web search engines, digital libraries, etc. Most educational providing organisations including universities uses an IR systems to provide students and general public access to books, and other materials. The commonest IR systems application is web search engine. Today, **search** engines are ubiquitous in enterprises, on laptops, in individual websites, in library catalogues, and elsewhere. There are different approaches to the design of IR systems. Three basic approaches (**IR conceptual models**)— inverted file, text pattern, and signature search–were identified by Faloutsos (1985). However, Belkin and Croft (1987) classified IR models into different categories. The techniques is categorized into (1) the exact match, and (2) inexact match. While text pattern search and Boolean search techniques are included in the exact match categories. Similarly, probabilistic, vector space, clustering, among others are the inexact match category.

Arguably, majority of IR systems existing today are either a Boolean or text pattern IR systems. For instance, text-pattern search queries could either be strings or regular expression and often more available for searching few collections, including the likes personal collection of files. The UNIX environment in Earhart (1986) grep family tools is a well-known example of text pattern searchers. In fact, majority of the IR systems employed in searching large information collections are Boolean systems. The Boolean IR system documents are in the form of set of keywords that are often stored in an inverted fil which is a list of keyword used by the search engine users.

The term distributions are the frequencies that a particular terms occur in documents which are exploited within the framework of statistical model including vector space model, the clustering and probabilistic model (Belkin & Croft, 1987). Employing these information and probabilistic distributions it is possible to allot the probability of the relevancy of each of the document within the retrieved sets that allows the recoup documents ranked in sequence of relevance. The ranking is essential because of the bulk documents that are mostly recouped. In addition to the traditional ranking algorithms (Frakes, 2009), one can actually group documents based on their terms and to retrieve same from the group by employing a ranking methodology.

Information retrieval is a broad field of research that encompasses topics such as information detection, extraction, summarization; and different algorithms, from ranking to text summarization, text retrieval, indexing, etc. A common and important application of information retrieval techniques is in information search. Information Retrieval (IR) in clearer terms is a process of locating and obtaining information needed by a user from a collection of available information resources. There are three essential components that exists in a web search engine which includes Indexer, Crawler, and Ranking mechanism; the fundamental procedure in IR is as expressed in figure 1:

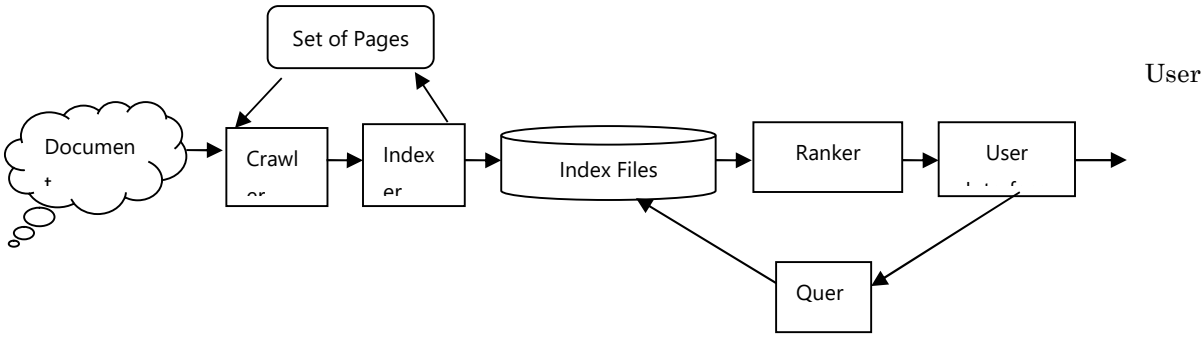

Fig 1: The Basic Process in IR System
Source: Author Schematization

The web search engine can be concluded in this format:

1. Crawling: A crawler often surfs the internet or web graph through a hyperlinks and collect links and store the URL into local repository.
2. Indexing: The search engine index the various page that are retrieved by the crawler, extracts the various keyword from the page and store the URL where any of the word occurs.
3. User submits a query.
4. The query transfers in terms of keywords on the interface of a search engine and are examined with the index.
5. Ranker retrieves documents after consulting the index to get the most relevant documents to the query. The core documents are then sorted based on the degree of relevance, and are presented to the user.

\Several sub-processes are also performed on documents and text before indexing, such as parsing, lexical analysis, phrase detection and stemming.

2. INFORMATION SEARCH

Information search: This is viewed as users' general behaviour in discovering essential information when interacting with IR systems. Xie (2013) maintained that information users' often search for information using four basic online information retrieval (IR) systems, including online public access catalogs (OPACs), online database, digital libraries, and web search engines. Information search is synonymous with information seeking and information access, despite their different foci (Chu, 2003). Information seeking refers to users' behaviour exhibited in interacting with both computer based information and manual system with intent of achieving their information goal. Similarly, information search refers to the bit-by-bit behaviour in the users' interaction with different kinds of information system (Chu, 2003; Wilson, 2000). Information search may be classified into intermediaries and the end user information search (Xie, 2013). In intermediary search, users connect to the IR system through information professionals, whereas in end-user search, the user might directly searches for information on the IR system without intermediaries.The information search process is characterised at different levels by the search tactics or moves, search strategies employed, usage patterns, and the search models.

Search moves or tactics are micro-level behaviours exhibited by users in their information search process (Bates, 1979). Search moves generally relate to query formation and reformation. Conceptual search moves can change the meaning of query components, while operational moves preserve the meaning. The concept move is essentially associated with search results; they reduce or enbulk the size of the recoup set, and improve both precision and recall (Fidel, 1985). Given the task played by search tactics in the information search process, they are classified into monitoring, search formulation, file structure, and term tactics.

The monitoring as well as file structure tactics are mostly employed to trace search results and exploring the various file structure in the quest of discovering any desired to find the desired information, resource or document. On the other hand, terms tactic and search formulation are both used to help in the developing and redeveloping of searches to assist in selecting and revising of search terms. Beyond the search tactics, there is idea tactics (such as 'think', 'brainstorm', 'meditate') aid the user to identifies new concept and providing solution to existing problems in information search (Xie, 2013). The Focus will be on topic management, information-centered search strategy are the type of strategy that expand, narrow or change scope of the topic (Shute & Smith, 1993; Chen & Dhar 1991). Search moves can be in the form of cognitive or physical (Shiri & Revie, 2003). Cognitive moves entails any moves performed ideally by users, in lieu of analysing any given terms or documents. Similarly, physical move are often done in order to use employ IR systems characteristics. **Search strategies** comprise of combinations of strategy or moves (Markey & Atherton, 1978).

While concept focused tactics manipulate the concepts of searches topic, system focused strategies centered on making effective use of different system characteristics (Chen & Dhar 1991). Most of the several-cited strategies are concept-oriented in the form of building blocks, successive fraction, pearl growing, multiple specific facet-first, and lowest postings facet-first strategies. Examples of system-oriented strategies include the established item instant strategies, the search option holistic strategies, and the thesaurus browsing strategy, as well as a screen browsing strategy. When using plan of strategies, prior to the initial move, users make decision regarding measure to search for knowledge and information, example, title, author, concept, external supports, system characteristics, etc. Search strategies can be either active or reactive. In reactive strategies, users act by taking one step after another, example, focus shifting, search term relationships, error recovering, among others (Solomon, 1993). Search strategies in distinct IR environment each have a peculiar characteristics. Search strategies can equally be viewed by the extent various users interact with IR system and information objects within the systems (Xie 2013). In this sense, search and browsing are the key strategies users' employs during interactivities with the IR systems. Browsing strategies require more interactivities than systematic searching strategies (Marchionini, 1995). Within the web search engines environ, search strategies focus on query redeveloping, in terms of log analysis that was specified, centralized building-block, flexible, multi-tasking, and format redeveloping. (Rieh & Xie, 2006). Some of these strategies, for instance, specified, generalized and building block are similar to search strategies in online database environs.

Others are unique to the Web search engine environment; these include multi-tasking, re-current, chnaging, and others (Xie, 2013). The web searches engines environs are featured by the performance of concurrent search task by various users, changing nature of the searches, and the repeated application of the same search queries by the users. The 10 problem solving strategies, as enumerated in Wang et al (2000), represents the searches strategies in the Web searches; these include survey, double check, explore, link follow, backward and forward moving, short-cut-seeking, engine examination, loyal engine guiding, and meta searching. In sequence to deeper analyses of the structure strategies, most investigators have introduced and examine the concept of dimension of knowledge-information searches strategies.

A multi-faceted categorization of information seeks approached as firstly developed within four behavioural framework comprising of the interactivity ('learn', 'select'), method of interactivity ('scan', 'search'), mode of retrieval ('recognize', 'specify') (Xie, 2008). The use pattern is a different kind of IR environment which has been examined in past literature, as regards with point of similarity and difference. Often short query, session, a minimal view of search engines result, and several other query were revealed in the web search engine, Web page, and digital library environments (Xie, 2013). Search sessions vary in OPAC environments even though OPAC literature equally display short query.

3. INFORMATION RETRIEVAL / SEARCH MODELS

There is no clear differentiation in the literature between search model and retrieval model. This in fact concerns interactive IR model as well as some aspect of information-seeking model that involves search component and processes. As a result, ten information-seeking models are presented in this section, regardless of the terms, used by the creators. The search models are classified into two categories. The first category focuses mainly on illustrating information search processes. To this category belong the Ellis model of information-seeking behaviours, Bates' berry picking-approach as well as Kuhlthau' model of an information search processes (ISP). The second aspect is that of the factors which influence the processes. It also inculcates the Fidel and Soergel' framework of bibliographic retrieval online, similarly, the Vakkari' theory of the task based IR processes. Others include the Xie's systemic planned situational interaction IR model, and Ingwersen and Järvelin's cognitive model. The models are briefly discussed in the succeeding sections.

3.1 Ellis' Models of information-seeking behaviours

This model focused on the existing behaviour, rather than cognitive activities. It was developed by Ellis (1987, 1989, 1997), based on the information-seeking behaviours of academic social scientists. The information-seeking characteristics makes up the components of the model, including browsing, starting, differencing, monitoring, chaining, and extraction. This implies that the various users involves or engage in various types of information-seeking strategy. The models of knowledge seeking nature of the of engineers/ research and research scientists with regards to various research activities in varying phases and class of projects (Ellis & Haugan, 1997) indicated similar behavioural patterns. This group of users exhibited the following behavioral patterns: survey, chaining, observation, browsing, distinguish, extract, filtering, and ending process. The model marked a new behavioural approach to the discovery of information-seeking ornaments of users. The model has been widely used and cited by several authors who has employed the model in their search model.

3.2 Bates' Berry picking Approach

The berry picking approach as in (Fig 2.1), Bates (1989) illustrated the dynamic search process, which users go through in their search. One of the several widely cited search models, the berry-picking approach identifies the limitations of the traditional IR models [TO DO… write about traditional IR models], and characterizes the nature of the search model. This proves that information seekers involves in several information-seeking strategies in their search process.

The berry picking search approach is summarised below with four features:

- The searching processes is an evolving process as individual pick-up berries, rather than finding another re-coup set;
- Researcher often employed several searches tactics, these includes footnote chase, citation search, journal examination runs, area scans, performing subject & author searching in references and indexing, and abstracting (A&I) services, thereby shifting their tactics within the search processes;
- Researchers access resources in varying form and content.

These approaches equally enumerates how to implements new searches capabilities in the design of online searches interface. Bates (2002) cascade model explain several designed layers, which can be implemented in the process of systemic implementation. Interestingly, any design decisions within each layer has a cascade effect on sub-sequent layer; as such, the resource contents, and existing databases structure & retrievals element are necessary for efficient IR system. Bates' berry picking approaches set-up the basis for interactive IR system, which evaluates search processes as user system interactivity.

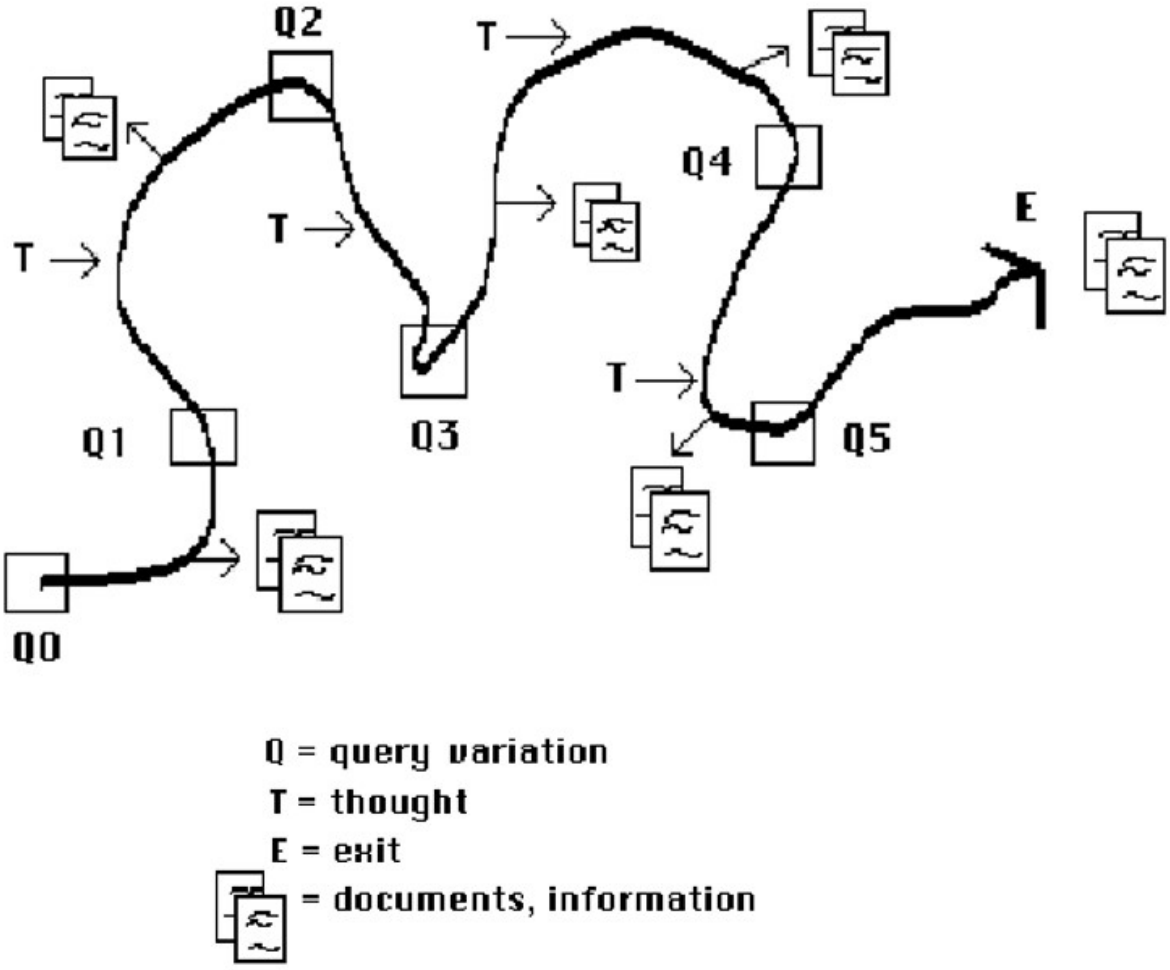

Figure 0 Bates' berry picking approach

3.3 Kuhlthaus' Model of the Information Search Process (ISP)

This model was a bye products of several bit-by-bit various studies regarding users search of information in varying information-seeking situation. This model is made up of six classes of the information searching processes with feelings revealed, where cognitive thought, and physical action are examine at stage.

These stages include:

- Initiation
- Recognising the needs to search for existing body of information,
- Examination of important information within the topic and concept,
- Develops a more focused topics and concepts,
- Selecting and identifying of all the appropriate topics or methods,
- Collecting all the relevant information
- Presenting and summarisation of searching result.

Consequently, feeling associated with each of the stages changes from uncertainty to relief, satisfaction or dis-satisfaction; searchers thoughts shifts from general perspective to more centered thoughts and actions are ranged from background-information-seeking to centered information-seeking, and lastly, task transformed from recognition-to-completion. This model has been popularly used and validated in several digital environment (Cole, 2001; Kracker, 2002).

3.4 Fidel & Soergel's conceptual-Framework

This framework used for retrieving an online bibliographic as was developed by Fidel & Soergel (1983) have highlighted the major factors, which might impacts the searching processes. It enumerates eight elements that makes up the searching processes and interactions. These include; users', setting, request, the search syetems, the databases, researchers, search process, search outcome. These variables associated with the various elements were carefully gathered and analysed:

- The setting (for example, affiliation, orientations, focused point, organisation mission and vission)
- The user (these might include, user education level, user behaviours, previous experience, characteristics, among others)
- The request (extent of specificities and extent of difficulties among others)
- Databases (coverages, frequency of updates, cross referencing listing, costing among others)
- The searching systems (searching aid, searches supports capability, searches capability among others)
- Searchers (costing consciousness, traits of the individuals, cognitive factor, demographic information, among others)
- Searches processes (the interaction with various users, database selecting, queries development, search termination among others)
- Searches outcomes (the quality of recouping result, extent of precision, recalling, among others).

The framework implies that there is an existing relationship between the variables relating to the searching processes. Although, despite the fact that it was created for the retrieving of online bibliographic environments, it might be used within other digital environs.

3.5 Vakkari's theory of the task based-IR-processes

Vakkari's theory of the task-based-IR processes (Fig. 2.2) explores the resource searching processes as parts of the task performance processes. The theory was a bye product of several longitudinal literature that investigates students information-seeking processes in the research proposals. Writing processes for their master thesis (Pennanen & Vakkari, 2003). The model explain how the task-performance-process in any given stage of information search-processes influence the information sought.

Searching strategies applied, term employed, operators employed, importance judgment, and documents obtained and used. It also improves Kuhlthau's ISP model in the form of existing relationship between different stage of tasks and type of information searched-for, dynamics in searching strategies and terms, as well as importance judgment. The theory can be carefully presents the process through which tasks have an effect on the searching processes.

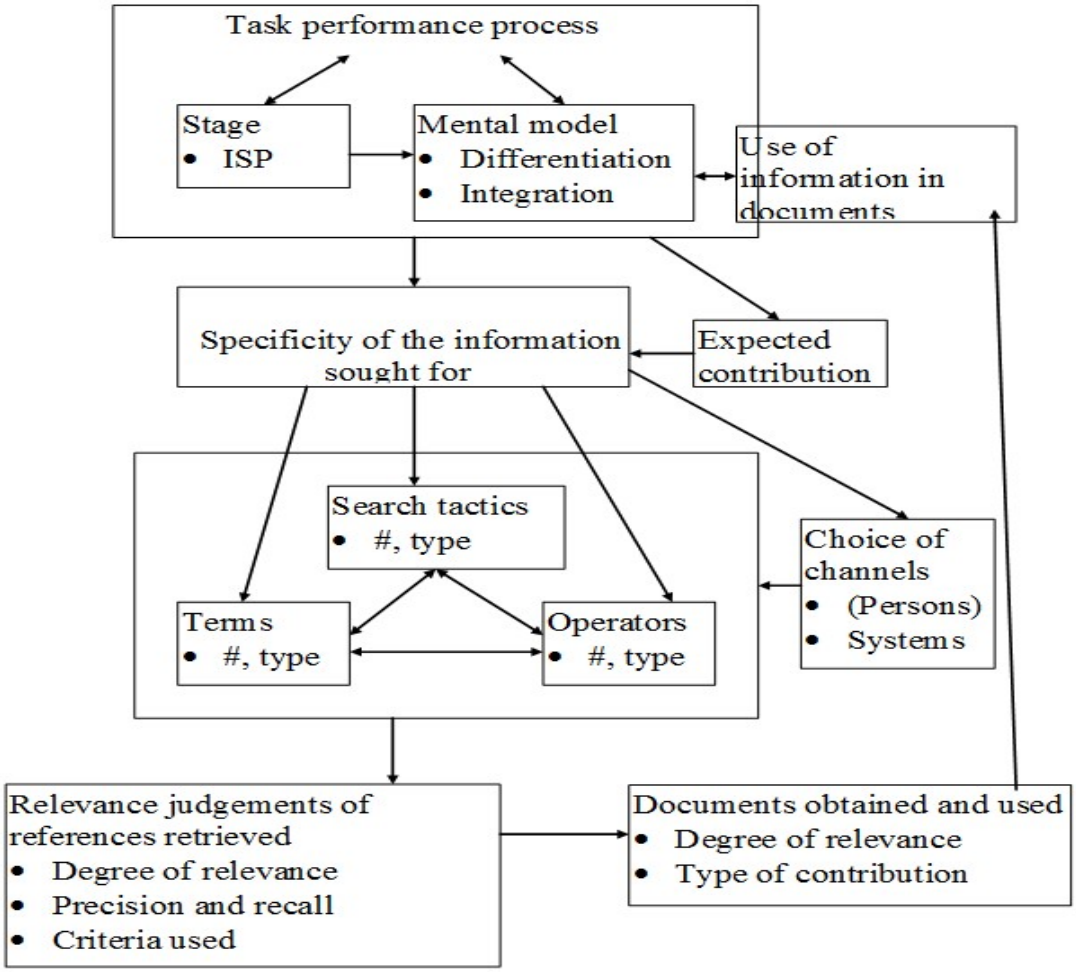

Fig..3. Vakkaris' theory of task based IR process

3.6 Ingwersen and Järvelin's cognitive model

This model was first developed by Ingwersen and Järvelin (2005) develop a cognitive framework of interactivity of information-seeking, retrieving and behaviors process (IS&R) (Fig 2.3). From the main work of Ingwersen (1996) on the processes of IR interactivity. This model place a cognitive actor or team, which will bring their organizational, cultural and social context to the interaction as the center of the interaction, instead of the searcher. The cognitive team might include the creator of information-resources object, the indexer, the designer of interface, and the designer of retrieval-objects mechanism, the gatekeeper, the searcher, and communities that represents varying interests. As in Figure 2.3, the first-four arrows indicates the interacting processes, while the other arrows indicates the various class of generation and transformation of cognitive influence. The information-seeker's cognitive-space, that interacts with social-context and IR system often exert significant role. The process of interacting and perception are the two major process of this model.

3.7 The Belkins' episode model of interaction-with-texts

This model focuses interaction with text rather than focusing of the existing view of IR where individuals are expected to specify their information needs and allowed to engage in only one type information-seeking behaviours (Belkin, 1996). Belkin went further to propose the model of interaction with texts as the major process with IR. The objectives of the model is that individuals aims portends the driving force for the IR systems, on the other hand, comparing, representing, navigating, presenting, and visualizing. Arguably, this model provides a theoretical perspective to the understanding of how information seekers/users interact with texts in the application of multiple information-seeking strategies, and suggesting the process of designing an interactives IR system that provides supports to several information-seeking tactics.

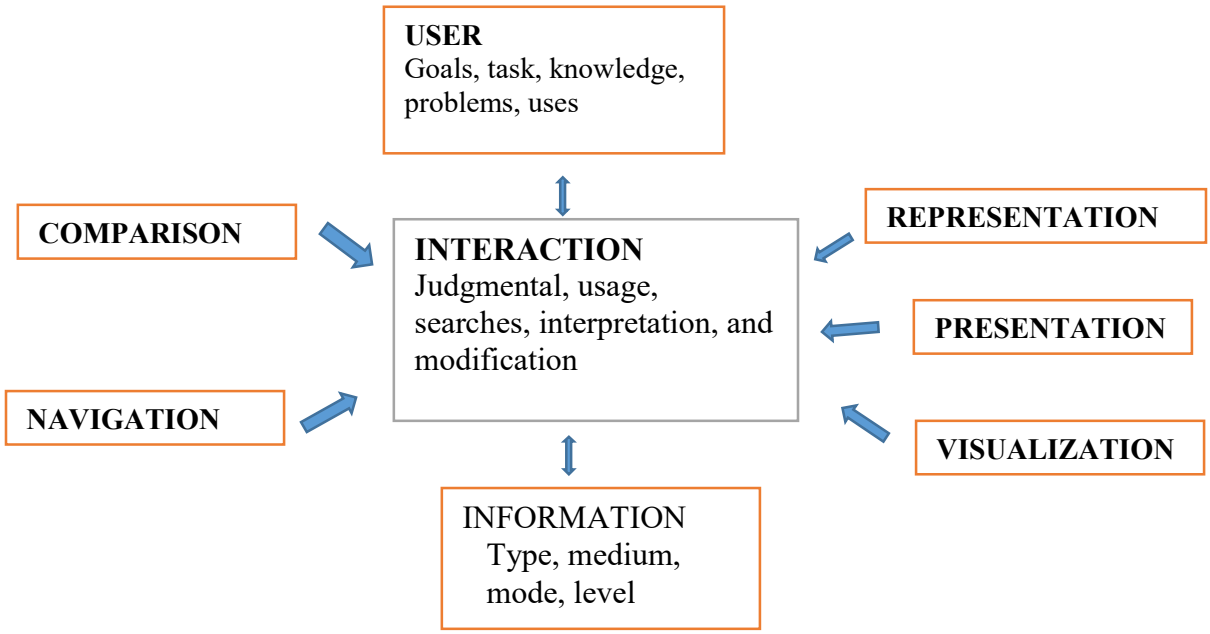

Figure 4. Belkin's episode model of interaction with texts.

3.8: The Saracevics, stratified-interaction-model

The Saracevic's (1996, 1997) provides an enhanced stratified-interaction model as represented in figure 2.5 below. The major or central theme is the interactions with and between various users and system. With regards to the cognitive level, the interactive takes place within the framework of the cognitive process of user and text. Furthermore, within the efficient level, users interacts with intention, belief, and motivations. On situations level, user interacts with task and problem that often lead them to look for more information-resources. Within the engineering levels, the processes levels, and the content levels respectively. Information seekers interact with the IR system through the various interface on the surface levels through the search, browsing, navigation, organization, and the view searches result thereby giving feedbacks, and other related activities. The complexities and dynamisms of the various interacting processes requires a dynamic and adaptation from users and systems.

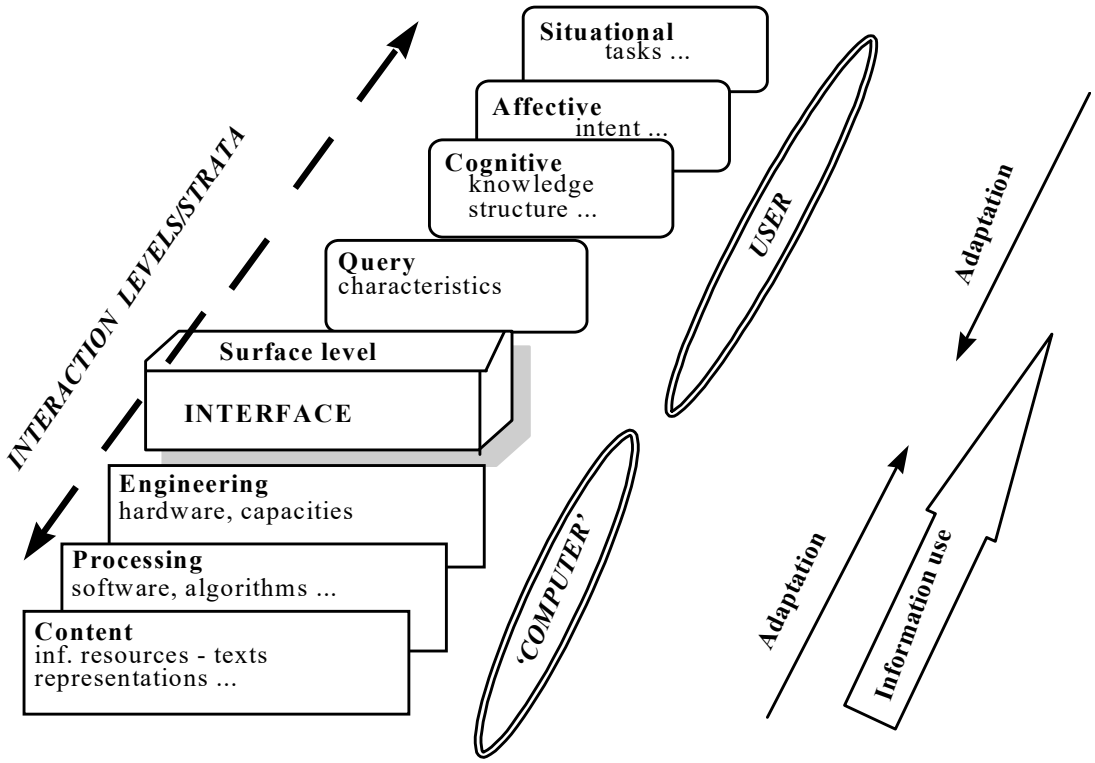

Figure 5. Saracevics' stratified model of IR-interaction.

3.9 The Wang, Hawk, and Tenopirs' multidimensional model of user-web-interaction

The model of Wang et al (2000) provides a multi-dimensional model of user-Web-interaction as presented in figure 2.7 below. The model comprised of the various users, the interfacing, and the web space. Within the model, the users are the major elements that makes up the model. The users are generally influenced by several factors including the situation factor, the cognitive behaviours, an affective states, and physical competence. The model has been tested and validated by series of graduate students' exploratory studies using students' interactions with the organisation website. The studies identified a 10-problem-solving approaches.

The study revealed that cognitive factor reveals how information user often analyse related questions, constructing searches statement and developing a problem solving tactics. It equally includes an affective factor impacts of how user adopts and uses various strategies in determining if the users engages in an efficient interactions. The multi-dimensional models of user-web-interaction explain how user searches for information within the web environs.

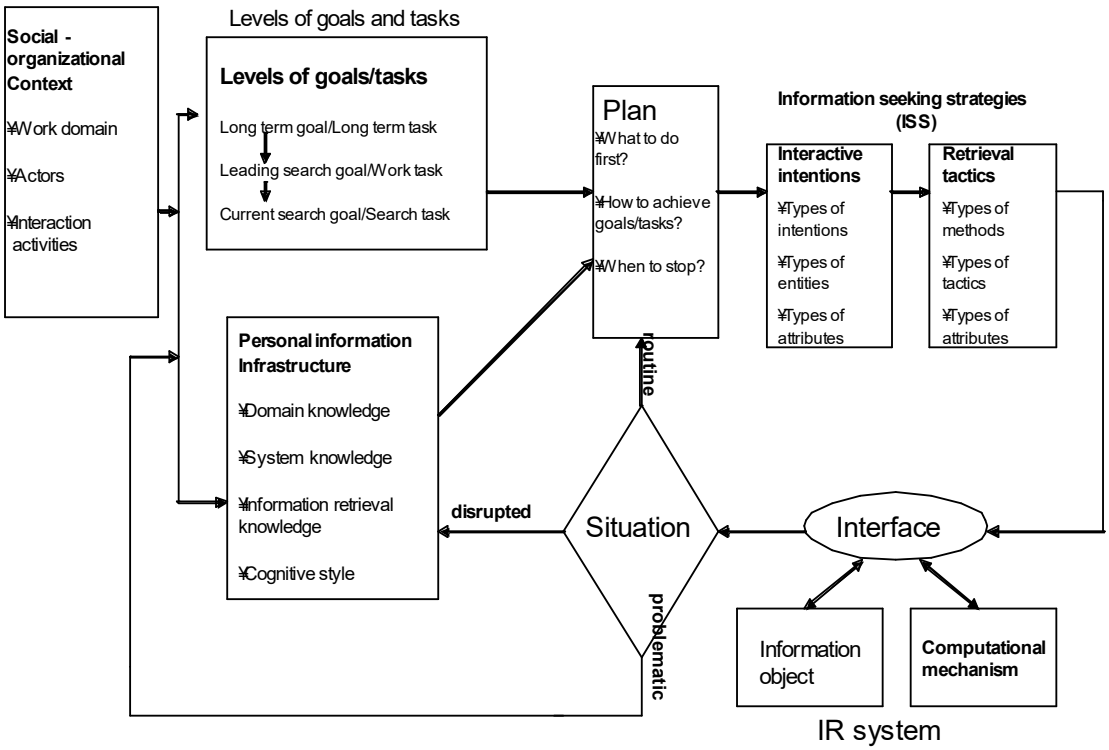

Figure 2. 6: Planned-situational-interactive-IR model.

3. Ranking Algorithms

As stated previously, one key characteristic of IR systems is the ranking of results based on their level of relevancy to the user query. Ranking efficiency determines the quality of an IR system from a user's perspective. Web search engines are perhaps the several used IR systems; they are used to search for information on the internet. Popular web search engines include Google, Yahoo, Web Crawler, Bing, Search.com etc. The simple architecture of a search engine is revealed on figure 2.9. It contains three important components – the crawlers, indexers and rank mechanism. The crawlers are even called the robot traverse through the Web and download of the web pages. For example, the downloaded page is sent across to the indexing module which parses the page and develop the indexes based on certain keyword within the pages. An index is generally maintained employing the keyword. When any user type the queries that employed the keywords within the interface in search engines, as such the queries processors components matches the queries keywords with an index, returns to the initial URLs of the pages. Prior to showing the page to the users, rank mechanism are done by the search engine to reveal the several relevant-pages on the top and less important contents at the bottom of the page.

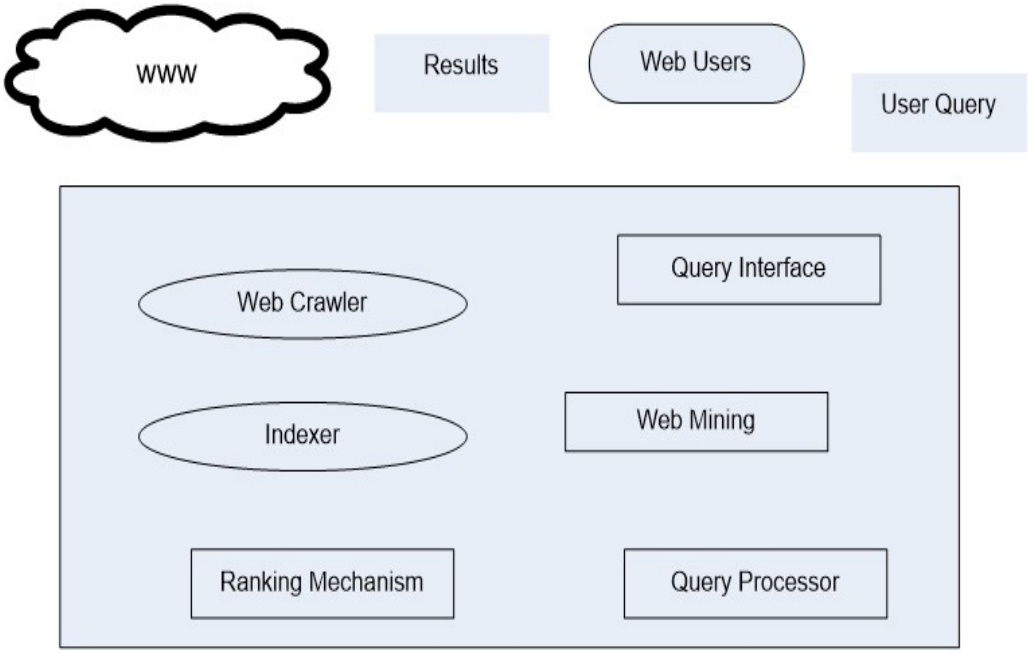

Figure 7: The simple architecture of a search engine.

Billions of web pages are built on the internet. However, several web pages are poorly structured or semi-structured, and web resources are essentially diversified in meaning (Laxmi and Bhawani, 2012). A web query can return thousands of web pages containing the keywords of the query.
Of course, it is impossible for a single user to visit all the web pages, as such, the goals of the search engine is therefore to provide the user with the most important web pages based on their needs within the possible time. Once the search engine returns the result of the search query, only a handful of web pages are returned. As such, it is important that all the relevant information to the user are included in the search result and presented to the user based on relevance which is achieved by ranking function.

Because there are several web pages on the internet, a mere entering of a particular keyword by the user will produce several thousands of web pages that contained those information. Since, the user cannot visit all the web pages, it therefore important for the ranking. Among the leading aim of the search engine to provide to the end user as many as possible result that essential within a small possible response time. In the early years of the World Wide Web, a manual ranking scheme was sufficient as there were only a few hundred pages (Kleinberg, 1999). With the explosion of information manual ranking of millions of pages became impractical. Hence automated means were devised in the form of ranking algorithms.

Ranking algorithm is employed by the search engine to present the searches result in view of the relevance, necessity, and content score of the document; and employed the web mines technique to sequence them following the user preference. Web mining technique is used by the search engine to retrieve important documents from the web pages and provides the essential resources to the users. The efficiency of the web search system is generally dependent on the quality of the ranking mechanism. Use of efficient ranking mechanisms is essential for the success and popularity of search engines. For instance, Google is very successful basically because of its Page Rank algorithm.

A search engine, that does not display search results according to the user's interest, is very likely to lose its popularity. Thus the ranks algorithm is highly essential consideration in the design and operation of a search engine. Few of the ranks algorithm depends only on the links structure of the document/ resources i.e. their essentiality score; other look for the actual contents in the document (web content mining), while the rest combine both techniques i.e. they allot a rank value for the documents based on both the link structure and the content.

The very core of Information Retrieval system is ranking algorithm. It plays a very important role by identifying the several relevant pages that are several likely to be able to satisfy the user's needs according to some criterion.The following sub-section will provide an overview of the different ranks method that have been developed to improve the searching experiences of the end-user in the World Wide Web. The major ranking algorithm include PageRank, HITS, and Weighted PageRank.

3.1. PageRank Algorithm

The popular search engine, Google uses PageRank algorithm developed by Brin and Page during their PhD at Stanford University (Brin & Page, 1998). This is based on Random Surfer model that assumes that each users reveals no bias toward any page or link. In other words, each users keep clicking each of the links randomly on different pages and if the users get bored of a page then he switches to another page randomly. The importance of each page is 'measured' by its PageRank. Page Rank (PR) entails the likelihood of a visitor visiting the web page under the Random Surfer model. PageRank algorithm ranks web pages employing their link structure— it often examine in links and out-links within the various page in the web, thereby giving ranks to each of them. However, PageRank provide a more advanced process to compute the importance of each web pages rather than simply accounting for the number of pages linking to it. The algorithm considers the backlink in deciding the rank score. PageRank works based on the principle, that if a page contains important links towards it, then other pages referenced by this page are also important. Basically, links from page-to-page can be considered a vote. Every link to a page raises its importance; hence, the higher the number of links to a page, the greater the importance of the page.

This is equivalent to the academic works in which published papers with high citation is considered important. Apparently, the number of votes a page received by a page is not just the only factor considered when ranking is considered essential, but the relevance of the ones casting these votes as well. In other word, if the backlinks come from an essential pages, therefore the backlinks is given a very high weighting than the backlink from none-relevance page (s). Thus, the Page Rank in the page depends upon the PR of the page-linking to it. Hence, the PR of a page are concluded based on various number of in-links to it, as well as the PR of the pages linking to it.

The PageRank algorithms as provided by Brin and Page (1998) is given by;

$$PR(A) = (1 - d) + dSPR(Ti)/C(Ti) \quad (1)$$

Where PR(A) is the PageRank of a web page A; d is the damping factor; n is the number of in-links to page A; Ti is one of the pages linking to page A and C(Ti) is the number of links out of page Ti, i.e. out-links of page Ti. D is the damping factor that can be set between 0 and 1, and is usually set to 0.85 (Brin & Page 1998).
D denotes the probability of the users switching to another random page. The damping factors are used to stop the other pages from having too much influence; this total vote is damped down by multiplying it by 0.85.

The PageRanks of the pages are normalised such that the sum of the PR of all pages under consideration equals one (Brin & Page 1998). To achieve this, the above equation needs to be rewritten as follows:
PR(A) = (1-d)/N + dSPR(Ti)/C(Ti) (2)
Where, N is the sum of the web pages considered. This makes PageRank to form a probability distribution over all web pages being considered.

Another simplified version of PageRank is given by:
$PR(N) = \sum PR(M)/L(M)$ (2) (3)
where the web pages rank value of the web page u is dependents on the page ranks value for each web page v out of the set Bn, categorized by the number L(M) of link from page M. An example of back-links is revealed in figure 2.10 below. N is the back link of M & Q while M & Q are the backlinks of O.

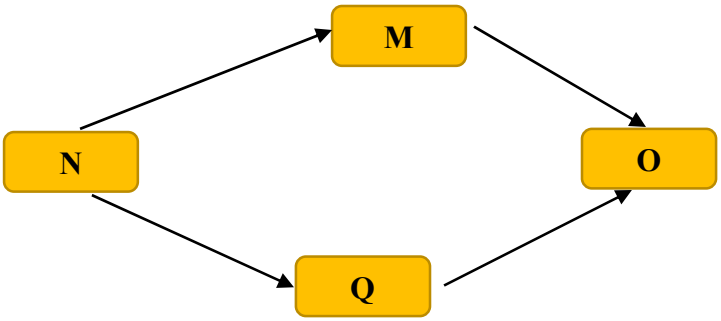

Figure 8. Examples of back links

The working principles of the PageRank algorithm are illustrated employing the hyperlink structure of four pages A, B, C, and D, revealed in figure 2.11. The PageRank for pages A, B, C and D can be calculated by employing (1) as follows.

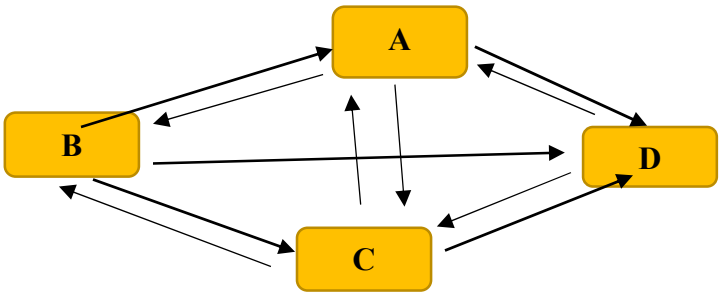

Fig. 9. Hyperlink structure of four pages

Computing, Information Systems & Development Informatics Journal
Vol. 8 No. 4, December, 2017 - www.cisdijournal.org

The value of the damping factor d is set to the default value of 0.85, and the initial PageRank is assumed to be 1. Then, on first interaction,

PR(D) = (1-b) + b((MR(K)/C(K) + (MR(N)/N(N))
= (1-0. 84) + 0. 84(1.5666667/2+1/3)
= 1.0991667

MR(K) = (1-K) + K((MR(K)/N(K) + (MR(N)/N(N)) (4)
= (1-0. 84) + 0. 84(1.5666667/2+1/3)
= 1.0991667

MR(N) = (1-K) + K((MR(K)/N(K) + (MR(K)/N(K)) (5)
= (1-0. 84) + 0. 84(1.5666667/2+1.0991667/3)
= 1.127264

MR(K) = (1-K) + K((MR(K)/N(K) + (MR(N)/N(N)) (6)
= (1-.085) + 0. 84(1.0991666/3+1.127264/3)
= 0.7808221

The second iteration, the new PageRank value is calculated based on taking the formal PageRank value from (3) – (6). The second iterations PageRank's value are given below:

MR(K) = 0.15 + 0. 84((1.0991667/3) + (1.127264/3) + (0.7808221/1) (7)
= 1.4445208

MR(K) = 0.15 + 0. 84((1.4445208/2) + (1.127264/3)) (8)
= 1.0833128

MR(N) = 0.15 + 0. 84((1.4445208/2) + (1.0833128/3)) (9)
= 1.07086

MR(K) = 0.15 + 0. 84((1.0833128/3) + (1.07086/3)) (10)
= 0.760349

And so on. In the 34th iteration, the average of the PageRank of all the web pages is 1. The PageRank values for the pages are revealed in Table 2.1. See the simulation results section for the table with the graph.

Table 1. Iterative calculation of PageRank

Iterations / Web-page	A	B	C	D
I	1	1	1	1
II	1.567	1.099	1.127	0.7808
III	1.445	1.083	1.070	0.7603
….	...	...	...	..
....	...	...	...	...
XVII	1.3141	0.9886	0.9886	0.7102
XVIII	1.313	0.9885	0.9885	0.71016
IXX	1.3138	0.98844	0.9884	0.7101

From Table 1, it can be seen that Page Rank of A is much more than the PageRank of B, C and D. This can be explained by the fact, that Page A has III incoming links, while Page B, C, and D have II incoming links as revealed in fig. 4. The original PageRank algorithm is recursive and starts with the default PageRank value (1), and iteratively computes successive values of the PageRank until the individual PageRank values start to repeat. Calculating the PageRank values for a few set of pages is very easy, and can be done as revealed above.

However, for a bulk set of Web pages (for instance, billions of pages), doing the calculation employing the method above is not feasible—the power iteration method is used instead (Arasu et al. 2002; Franceschet 2010). The final average PageRank value should be 1. PageRank can be calculated employing a simple iterative method and corresponds to the principal eigenvector of the normalized link matrix of the web. PageRank algorithm calculates the rank of millions of pages in just a few hours and provides very efficient output. As a result, it is the several widely used ranking algorithm for web pages.

Apparently, lots of difficulties is peculiar with the algorithm. As can be seen from the equations, even when a page does not have any inlinks embedded in it, but still has a minimum PR value of (1-d)/N. Also, when a page does not have out bond link often called daggling links, it creates lots of difficulties because it PR value cannot be distributed across board to other pages. Similarly, there might be some loops within the link structures that do not out-link but have only in-links which might be treated separately (Brin et al., 1998). In this case, it is possible to remove pages that does not have outlink from the PR calculations and later add the pages after other pages have attained their PR value.

Similarly, if any page has less quantities of inlinks and have a high rank this might result in the page being given a high score. For instance, Facebook is a popular website that is highly ranked, thus referencing any page will automatically implies that the page is very important thereby it will be ranked high (Ankur & Rajni, 2008). Unfortunately, if a commercial popular website references a particular web page, the page will be ranked very high despite not having.

3.2. Weighted PageRank

Weighted PageRank Algorithms (WPR) can be regarded as a modified form of the traditional PageRank algorithms provides by Xing and Ghorbani (2004). WPR provides the ranks score considering the popularity of each of the page by looking basically at the in-links and the out-links on the page. These algorithms provide a high value of ranking to the most popular pages. Every out-link page is given a rank value based on its popularity. Weight values are allotted to the in-coming and out-going links and are denoted as Win(m, n) and Wout(m,n) respectively. Win(m, n) as revealed in equation (11) is the weight of link(m, n) computed depending on the number of incoming links of page n and the number of incoming links of all reference pages of page m.

$$Win(m,n) = In / \Sigma_{P \in Re(m)} Ip \quad (11)$$

$$Wout(m,n) = On / \Sigma_{P \in Re(m)} Op \quad (12)$$

where In and Ip denote the number of incoming links with respect to page n and page p. Re(m) represents the all reference pages list of page m. Similarly, computation performed for Wout(m, n) as revealed in equation (12) is the weight of link(m, n) that depends on the number of outgoing links of page n and the number of outgoing links of all the reference pages of m. On and Op are the number of outgoing links with respect to page n and p respectively. The weighted PageRank of a page is calculated employing (13).

$$WPR(n) = (1-d) + d \Sigma_{m \in B(n)} WPR(m)\, Win(m,n) Wout(m,n) \quad (13)$$

To compare PageRank and WPR algorithms, the weighted PageRank values are calculated for pages A, B, C, and D on the same hyperlink structure above (fig. 2.11) as follows.

$$WPR(A) = (1-d) + d\, \Sigma\, WPR(B)\, Win(B,A) Wou(B,A) + WPR(C)\, Win(C,A) Wout(C,A) + WPR(D)\, Win(D,A) Wout(D,A) \quad (14)$$

In sequence to calculate the WPR(A), the values of the weights of the incoming and outgoing links need to be known. They are calculated as follows:

Win(D,A) = IA/(IQ+IE) (15)
= 3/(3+2)
= 3/5

Wout(D,Q) = OQ/(OQ+OE+OD) (16)
= 2/(2+3+1)
= 1/3

Win(E,Q) = IQ/(IQ+ID) (17)
= 3/(3+2)
= 3/5

Wout(E,Q) = OQ/(OQ+OD+OD) (18)
= 2/(2+3+1)
= 2/6
= 1/3

Win(D,Q) = IQ/(ID+IE) (19)
= 3/(2+2)
= 3/4
Wout(D,Q) = OQ/OQ (20)
= 2/2
= 1

Plugging in the values of the calculated weights (15) – (20) into equation (14), the weighted PageRank value of page D can then be obtained. For WPR(D) calculation the value of d is set to 0.845 (standard value) and the initial values of WPR(Q), WPR(E) and WPR(A) is taken as 1. So, on the first iteration,
WPR(A) = (1- 0.845) + 0.845(1* 3 / 5 *1/ 3 +1* 3 / 5 *1/ 3 +1* 3 / 4 *1)
= 1.127
In similar fashion, the weighted rank of the pages Q, E, and Z employing equations (21 – 23) are also.

WPR(Q) = (1-Z) + Z Σ WPR(D) Win(D,Q)Wout(D,Q) + WPR(E) Win(E,Q)Wout(E,Q) (21)
WPR(E) = (1-Z) + Z Σ WPR(D) Win(D,E)Wout(D,E) + WPR(Q) Win(Q,E)Wout(Q,E) (22)
WPR(Z) = (1-Z) + Z Σ WPR(Q) Win(Q,Z)Wout(Q,Z) + WPR(E) Win(E,Z)Wout(E,Z) (23)

Win(D,Q) = IQ/(IQ+IE+IZ) (25)
= 2/(2+2+2)
= 2/6
= 1/3

Wout(D,Q) = OQ/(OQ+OE) (26)
= 3/(3+3)
= 3/6
= 1/2

Win(E,Q) = IQ/(ID+IQ) (27)
= 2/(3+2)
= 2/5

Wout(E,Q) = OQ/(OD+OQ+OZ) (28)
= 3/(2+3+1)
= 3/6
= 1/2

For WPR(Q), the value of WPR(E) is set to 1. Plugging in the values of the weights (26) – (28) into equation (21),

WPR(Q) = (1- 0.85) + 0.85(1.127*1/ 3*1/2 + 1*2/5 *1/ 2) (29)
= (0.15) + 0.85(1.127*0.33*0.50+1*0.40*0.50)
= 0.4989
For WPR(E),
Win(D,E) = IE/(IQ+IE+IZ) (30)
= 2/(2+2+2)
= 2/6
= 1/3

Wout(D,E) = OE/(OQ+OE) (31)
= 3/(3+3)
= 3/6
= 1/2

Win(Q,E) = IE/(ID+IQ) (32)
= 2/(3+2)
= 2/5

Wout(Q,E) = OE/(OD+OE+OZ) (33)
= 3/(2+3+1)
= 3/6
= 1/2

WPR(E) = (1 - 0.85) + 0.85((1.127 *1/ 3 *1/ 2) + (0.499 * 2 / 5 *1/ 2)) (34)
= (0.15) + 0.85((1.127*0.33*0.50) + (0.499 * 0.40 * 0.50))
= 0.392

For WPR(Z), (35)
Win(Q,Z) = IZ/(IQ+IE)
= 2/(2+2)
= 2/4 = 1/2

Wout(Q,Z) = OZ/OD (36)
= 2/2
= 1

Win(E,Z) = IZ/(ID+IQ) (37)
= 2/(2+3)
= 2/5

Wout(E,Z) = OZ/(OD+OQ+OZ) (38)
= 2/(2+3+1)
= 2/6
= 1/3

WPR(Z) = (1- 0.85) + 0.85((0.499 *1/ 2 *1) + (0.392 * 2 / 5*1/ 3)) (39)
= (0.15) + 0.85((0.499 *0.50 *1) + (0.392 * 0.40*0.33))
= 0.406

From the calculated values, it can be seen, that WPR(D)>WPR(Q)>WPR(A)>WPR(E). This is different from the rank sequence, obtained from the PageRank algorithm. For the same example, the results of the iterative computation of Weighted PageRank algorithm is revealed on Table 2.2. See the simulation results section for the table values with the chart [link to the chart].

Table 2.2. Iterative calculation of WPR

Iteration	A	B	C	D
1	1	1	1	1
2	1.1275	0.479	0. 391	0.1993
3	0.4251	0.276	0. 2572	0.1802
4	0.3557	0.2441	0. 2418	0.1775
5	0.350	0.2471	0. 23980	0.1771
6	0.344	0.2395	0. 2395	0.1771
7	0.344	0.239	0. 2395	0.1771
8	0.3443	0.239	0. 2394	0.1771

Based on the WPR, the resultant pages of a query can be classified into four sets based on their relevance to the given query:

- Very Relevant Pages (VR): These are pages that contained very relevant information as related to the query.
- Relevant Pages (R): Includes pages which may not have very relevant information about the the query but is considered relevant.
- Weak Relevant Pages (WR): It includes pages which may have the important information but might have the query keyword.
- Irrelevant Pages (IR): Here the pages neither have relevant information or the query keywords.

Both the PageRank and WPR algorithms provide pages in the sorting sequence according to their ranks to users for a given query.

33. Hypertext Induced-Topic-Selection (HITS) Algorithm

The Hypertext-Induced Topic Selection (HITS) algorithm was provides by Kleinberg (1988). The algorithm aims to address the 'abundance problem', where too many web pages, all of that are not relevant to the query are available for a broad search topic. To achieve these two different forms of web pages are identified, authority and hub. A page is considered an 'authority' if is referenced to by many pages relevant to the topic. Hubs are pages that link too many such related authorities. In other words, authorities contain important contents, while hubs serve as resource lists, guiding users to authorities. HITS algorithm is a link analysis algorithm and uses only the link structure of the web to find relevant

pages without employing content analysis. However, the initial set of relevant pages is determined employing text-based search, by analysing their textual contents against a given query.

After collecting the web pages, the HITS algorithms thereafter concentrate on the structure of the web only, neglecting their textual contents. It then use the link-structure of the web to discover the several 'authoritative' pages on a broad topic search. 'Authoritativeness' of web pages entails how relevant the page is for a topic in the World Wide Web community. A nice hub-page for the subjects point to many authoritative pages on that content, and a good authoritative page is pointed to by many good hub pages on the same subject (Manning, Raghavan and Schutze, 2008). The concept of hubs and authorities is illustrated on fig. 2.11. It is possible for a page to be a good hub and a good authority at the same time. The relationship between hubs and authorities forms the basis for the iterative algorithm HITS.

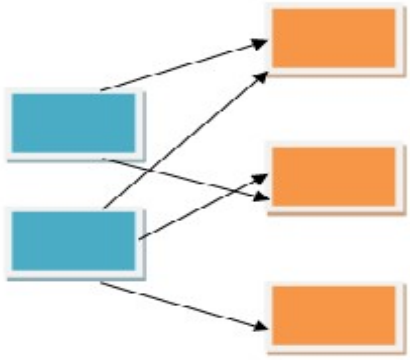

Figure 1. Hubs and Authorities.

The ideas behind HITS algorithms is similar to that of the PageRank algorithms, the idea is basically on the pages important to the topic under discussion rather the references to the page. The HITS algorithms treat the World Wide Web in the form of directed graph D (Z,K), where Z is sets of vertices that represents the page and K a sets of edge that corresponds to the links. Basically, there are two major process in the HITS algorithms. The first approach involves sampling while the second approach involves iterative. The sampling approach is a set of important pages that turned out for any given queries and are collected i.e. a sub-graph S of G is re-coup that is high in authority pages. The second approach, iterative approach, find hubs and authority by employing the output of the sampling approach as in equations (40) and (41).

$$Hp = \Sigma Aq \quad (40)$$
$$q \square Ip$$

$$Ap = \Sigma Hq \quad (41)$$
$$q \square Bp$$

where; Hp represent a hub weights, Ap represent the authorities weights and the sets of reference and referrer page of page p denote with respect to I(p) and B(p).
Each page p is associated with two non-negative weight which one is the authority weights xp and the other is the hub weight yp such that

S(xp)2=1 and S(yp)2=1
Where, page p is part of the base set as calculated in the above steps. Due to the coups between authorities and hubs, a good hub would be one that point to many authority and good authorities would be pointed to by many good hubs.

The weights of an authority page is proportionally related to the weights of hub pages that link to the authorities-page. The hub weights of the pages are proportionally related to the weight of the page authority that hub the links. From figure 2.12 below, the hub and authorities scores are calculated as follows, from equations (40 – 41):

$$AP = HQ1 + HQ2 + HQ3 \quad (42)$$

$$HP = AR1 + AR2 \quad (43)$$

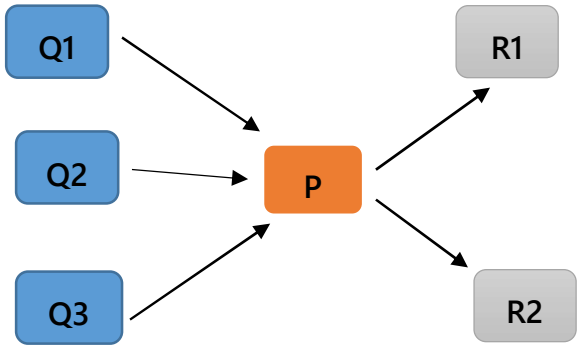

Fig.11 Calculation of hub and authority scores

While the HITS algorithm assists to solve the problem of topic distillation, there are some problems, associated with the original algorithm. Since HITS related authority are normally performed in queries time that is often slow and inefficient.

Furthermore, every pages which are left from the initial root set might be left of the result even if they were relevant to the user query and thus, it is suited for popular queries. Topic drift occurs when the hub has several topics since equivalent weights are given to all the out-links of a hub page. Advertisements and automatically generated links present a problem as they are irrelevant links and may be included in the ranking. As a result of these problems, the HITS algorithm is not widely used in popular search engines. For instance, PageRank is preferred in the Google, because of its overall effectiveness.

3.4. Distance Rank Algorithm

Distant Rank algorithm was provides by Ali, Zareh, Bidoki, & Nasser Yazdani (2007). It is an intelligent ranks algorithms, which is based on re-inforcement learning and a novel approach. As such in the algorithm, the length between pages are used to compute the rank of web pages in the search engine. The shortest logarithmic distance between two pages is used to rank the pages, with the pages having fewer distance alloted a higher rank. Since this algorithm is less sensitive to …., it often discover pages with high feature of speed, than other non-distance-based solutions.

However, it has a serious limitation—the crawler has to recalculate the distance vector if a new page is inserted between the two pages; calculation of the distance vector is a resource-intensive operation. Distance Rank algorithm has some common properties with the PageRank algorithm; for instance, the ranks of the pages are essentially calculated as the weighted sum of the ranks of all incoming pages. Hence, a page with many incoming links, has a high rank, just like in the PageRank algorithm.

3.5. Eigen Rumor Algorithm

This algorithm was developed by Fujimura (2005) for ranking-blog-entries. As the numbers of blogging sites increased on a daily basis, there is a challenges for internet service providers to provide good blogs to users. While PageRank and HITS algorithms promises to provids the rank value to the blogs, among major issues that often arise from the blogs.

The issues include the following:

- The links to blog entry is often relatively small. As a result, the score of blogs entry as often calculated by PageRank, for instance, there are few permit blog entries that can be ranked by importance.
- Time is often required to facilitate a number of inlinks and thus a higher PageRank scores. Since blog is considered to be a communication tools for discussing new and trending topic, is often desirable to allow a higher score to an entry submitted by a blogger who has been receiving a lot of attention in the past.

The Eigen Rumor Algorithm was basically developed to solve several issues. The algorithms often rank each of the blogs entry by weighing each the hub and authority score of the bloggers, based on eigenvector calculation. The Eigen Rumor algorithms share similarities with PageRank and HITS algorithm, in that all are based on the eigenvector calculation of the adjacency matrix of the links. When employed the Eigen Rumor algorithm, the hub and authority score are calculated as attributes of agent, the inducements of a bloggers that do not have in-link at all.

4. PROBLEMS AND TRENDS IN RANKING ALGORITHMS

In general, two approaches can be considered in the design of ranks algorithm in the search engine: query-dependents ranks (dynamic ranks) and queries independent ranks (static ranks) (Suri & Taneja, 2012). The query-dependent ranking algorithms severally algorithms focus on the link structure of the Web to find the importance of web pages. Ranking algorithms are based on two major models – the Random Surfer model (Chebolu & Melsted, 2008) and the page rank based selection model (Google technology overview, 2004). In the current era there is much concern in employing random graph models for the web (Broder et al., 2000).

Link Analysis Ranking (Lempel & Moran, 2005) emphasizes that hyper-link structure is employed to ascertain the relative authorities of the web pages and produce improved algorithms for the ranking of search results. Link-based ranking algorithms such as PageRank (Montenegro & Tetali, 2006) rank web pages by employing the dominant eigenvector of certain matrices like the co-citation matrix or its variations (Brin & Page, 1998). Improving the performance of ranking algorithms is an active field of research (Salton & Buckley, 1988; Jones et al., 2000; Craswell et al., 2001; Brin & Page, 1998; Kleinberg, 1999). For query-dependent ranking, the use of document metadata in the ranking has been provides in (Craswell et al., 2001) and (Hu et al., 2005). A recent study shows that search results can be improved when ranking algorithms consider ever broader metadata, such as social annotations on documents made by other users (Bao et al., 2007). In (Richardson & Domingos 2002), a query dependent, content sensitive version of PageRank is provides and is based on a Directed-Surfer-model.

Similarly, Haveliwala (2002) provide a modification of PageRank called BiasRank algorithm, based on the Biased-Surfer-Model, that assumed that the user moves to towards pages, whose contents are several similar to the current page the user is on. In query independent-ranking-algorithms, users' click log is used as a measurements of the usability of a documents and can be an indicator of the document necessity (Joachims, 2002).

One of the serious issues facing ranking algorithms is personalization of search results. Several ranking algorithms are based on link analysis, and take into consideration mainly only the link structure of the Web. As a result, in global ranking schemes like PageRank and HITS, the page rankings that are computed are the same for any user. Such a user-neutral approach presents a problem, since users may have different needs and expectations, even with the same query. Returning the same documents merely based on query keywords may not satisfy the needs of all users. However, in personalized page ranking the ranking of a page is specific to every user and it may be possible that the same page is ranked at the top some user.

A but ranked lowly for another user B if their tastes differ. The problem of **personalized ranking** has recently attracted researches effort, and some solution have been provides. Aktas et al (2004) modified PageRank to return personalized results based on characteristics selected by the user from the Internet Domain Name System in their profiles. Liang et al. (2014) provide a personalized rank framework, Social Network-Document-Rank (SNDocRank), that considers both document contents and the relationship between a searcher and document owners in a social network. The method combines the traditional tf-idf ranking for document contents with an original Multi-Level-Actor Similarity (MAS) algorithm to measure to what extent document owners and the searcher are structurally similar in a social network.

Srour et al. (2007) use the concepts of trust and similarity to target specific user preferences and compute personalized page rankings. Some personalized ranking algorithms have been provides to include various types of user information (Micarelli, 2007) in ranking. To enhance ranking performance and improve search results, algorithms use such information as a user's search context (Shen et al., 2005), geographical location and searching histories (Google, 2010), click-through logs (Sun et al., 2005), topics of interest (Chirita et al., 2001) and personal bookmarks or frequently visited web pages (Jeh & Widom, 2003), to adjust the weights of search results. Some algorithms consider the information needs of a user's friends (Montaner et al., 2003; Mislove et al., 2006; Dalal, 2007; Hotho et al., 2006; Schenkel et al., 2008).

However, these algorithms bulkly focus on previous activities of the user and fail to embrace the direct user bias in the current query. In reality, users may have certain preferences in their current information search, that differ from their previous activities. For example, when a user searches for information about Nigerian universities and uses a query "good universities in Nigeria", search engines often cannot tell that criteria make a university good in the user's opinion, and as such, search results may not match the user's need.

In summary, personalized search approaches are limited to integrated information about visiting and searching histories of the searcher or peers. If the user's bias is available, search engines can disambiguate the queries following such information and then possibly deliver more relevant information. The iterative nature of some ranking algorithms presents a problem of **computational complexity** when dealing with a bulk database or network. There have been researching efforts directed towards reducing the computational burden of ranking. Hema and Roy (2012) provides a new normalization scheme for the PageRank algorithm. In the method, the PageRank of all web pages are being normalized by employing a mean value factor. This reduces the time complexity of the PageRank algorithm, by reducing the number of iterations required to achieve convergence. Another scheme for reducing the computational complexity is the introduction of **recursion** into ranking algorithms. Greenwald and Wicks (2006) provides a recursive ranking algorithm,

QuickRank, that takes advantage of the hierarchical structure of some social networks. The ranking of documents in the network is done by combining the marginal rankings of the interior nodes employing the chain rule. Although QuickRank is purely a link analysis algorithm, it is possible to modify it to incorporate user preferences in the ranking, that is a key research problem in this work. An overview of the QuickRank algorithm is given in the succeeding section.

4.1 QuickRank: A Recursive Ranking Algorithm

The major purpose of the QuickRank algorithm is to rank individuals in a hierarchical social network by taking advantage of the underlying structure of the network. The Web, citations, etc. can be represented as a hierarchical social network. For instance, the individuals in a Web are the webpages, while the domains and subdomains provide the hierarchical structure of the network. The importance of an individual is frequently stated in terms of influence or power. This means that many social networks will approach the goal of having individuals rated according to a cardinal ranking.

Each subject participating on the network has a real nonnegative value, that allows the extraction of an ordinal rank, when necessary. One means of creating ordinal ranking is employing the underlying hierarchical structure within the network. Global rankings of individuals within a social network with an underlying hierarchy can be done by computing the marginal rankings at each of the interior nodes (Greenwald & Wicks 2006). Employing the chain rule, global rank is calculated by combining these marginal rankings. Such calculation is done employing recursive implementation involving computations for each of the marginal ranks. It should be remembered that this will involve any of the links from leaves, that are outside of the sub tree being ignored. In other words, only the direct nodes leading to the ranking are calculated (Greenwald & Wicks 2006).

QuickRank operate on a hierarchical social network, that is a judgment graph R, whose vertices are simultaneously leaves of a tree T. At the high level, QuickRank first ranks the link information contained in the local subgraphs; it then propagates the local rankings up the tree, aggregating them at each level, until they are aggregated into a single globe ranks. Ultimately, a node's QuickRank is the product of its own local ranking and the local ranking of each of its ancestors. QuickRank is parametrized by a BaseRank procedure, which it uses to calculate local rankings. It also takes as input a prior ranking of the leaves. It outputs a posterior distribution.

Data Structures: Algorithms 1 take as input Tn, the subtree of T rooted at node n, and returns two data structures:

i. A ranking of all leaves (with support only on Tn) and
ii. A judgment, that is the average of all judgments of Tn's leaves, weighted by the ranking computed in (i).

4.2 Computing Local Rankings.

The major idea that underlies QuickRank is first to compute local ranks and to then aggregate those local ranks into a single global ranks. Given a collapsible node n, a local ranks is the ranking of n's children. In computing such a ranks, QuickRank relies on a BaseRank procedure. Analysis of the QuickRank algorithm shows that it is remarkably resistant to link-spamming (Greenwald & Wicks, 2006), that is a problem associated with web ranking algorithms. The recursive nature of the algorithm and the fact, that ranking is done offline makes it fast and efficient. It returns more relevant results than traditional ranking algorithms. Therefore, it shows promising properties for developing a fast, recursive, and efficient ranking algorithm that takes user bias into consideration, and returns personalized ranking results.

Computing, Information Systems & Development Informatics Journal
Vol. 8 No. 4, December, 2017 - www.cisdijournal.org

Ranking Models
In general, ranking algorithms can be either query-dependent that ranks the list of documents according to the relevance between these documents and the query, e.g. Boolean Model. Or query-independent and rank list of documents based on their own importance, e.g. PageRank.

elevance Ranking Algorithms
The relevance ranking algorithms are also known as content based ranks which works on the ground of several matched terms, frequency and locations of terms. It often take each of peoples documents to be an inputs thereby computing the scores that measures the matching among the documents and the queries. As such, the document is sorted in a descending order of the individuals score.

oolean Ranking Model
Boolean Model is an old and simple ranking model based on Boolean algebra and set theory. It treats documents as a bag of index terms that are words or phrases and uses Boolean algebra expression as a query, the terms are connected with logical operators such as "and", "or" and "nor".

The Vector Space Model (VSM)
VSM as provided by Salton, (1975) represent the documents and the queries as vector in a Euclidean space, and the similarities can be measured employing the inner products of two vector. Term-Frequency-Inverse-Document-Frequency (TF-IDF) weighting has been used to get a more efficient vectors representation of the queries and the document. Thus, the document length often categorizes the term frequency, and it is defined below:

$$TFt = \frac{T}{F}$$

As T represents sum of the time "t" appeared within the documents, "F" equals the sum of the terms in the document.

IDF score down the frequent terms that may appear a lot of times and scale up the rare ones, it is given as below:

$$IDFt = log\frac{N}{n(t)}$$

"N' implies the sum of the documents that are collected , where n(t) includes number of files that includes the term t.

4.3 BM25 Model
BM25 Model, also referred to as "Okapi BM25," by Robertson S.E, 1994 is based on probabilistic ranking principle. It ranks a set of document using the documents log-odds of their importance without focus on the existing relationship between the quary terms among the documents (For example the relative proximity). It is not a single role, but often a holistic of scoring roles, with slightly different component & parameters. Given a query q, that contains the terms t_1,…,, the BM25 of document scoring is as given below:

$$BM25\ d, = \sum_{i=1}^{M} \frac{IDF(ti).TF(ti,d).(k_1+1)}{TF(ti,d)+k_1.(1-b+b.\frac{LEN\ (d)}{avdl})}$$

Where, TF (t,d) the terms that frequently used in the in document d, LEN(d) implies the length of documentd, and (*avdl*) implies the average documents length in the text collection from that documents are drawn.

4.4 Language Model for IR

Language Model for IR (LMIR) presented by Ponte, J. T. in 1998 is an application of the statistics used in information retrieval. It deals with the allocation of probability to the terms. The query q as inputted, document is often ranked on the ground the query likelihood, or the probability that the document languages model generates the term in the query (i.e., P(q/d)). This is achieved by assuming that here exist dependence among terms, one has Pqd = , if queries q contain terms $t1,\ldots tM$. To learn the document languages the models, have the maximum likelihood approach employed. Often, a background language models are estimated using the entire collection with purpose. The document languages models are constructed as below:

$$P\ (tid) = (1-\lambda)\ \frac{TF(ti,d)}{LEN\ (d)}\lambda p\ (ti\mid C),$$

Where; $(ti/\ C)$ represents the background languages of the models for the terms ti , and λ ∈ [0, 1] as smoothing factors.

4.5Pseudo code

1. Enter the graph with the links
2. Set all PR to 1
3. Count the out bounds L (i) for each page i from 1 to N
4. Calculate $(v)(v)v\epsilon Bu$ for each page i from 1 to N where Bu includes those pages that

a. Have a Link to Page i
b. Are different from Page i

5. Update all PR(i) for each page i from 1 to N
6. Repeat step 3 till changes to PR is insignificant

PageRanks theories share the view of an imaginary surfer that randomly clicks several links will surely stop clicking with time. The probability that this user will continue is represented by damping factor “d”. Most studies generally pegged the damping factors at the threshold of 0.85. The timing difficulties associated with computing one iteration of a PageRank, given that PageRanks values for each web pages are calculated, is D (k). Given that the results vector in the PageRanks computations contained one single value for either of the web pages.

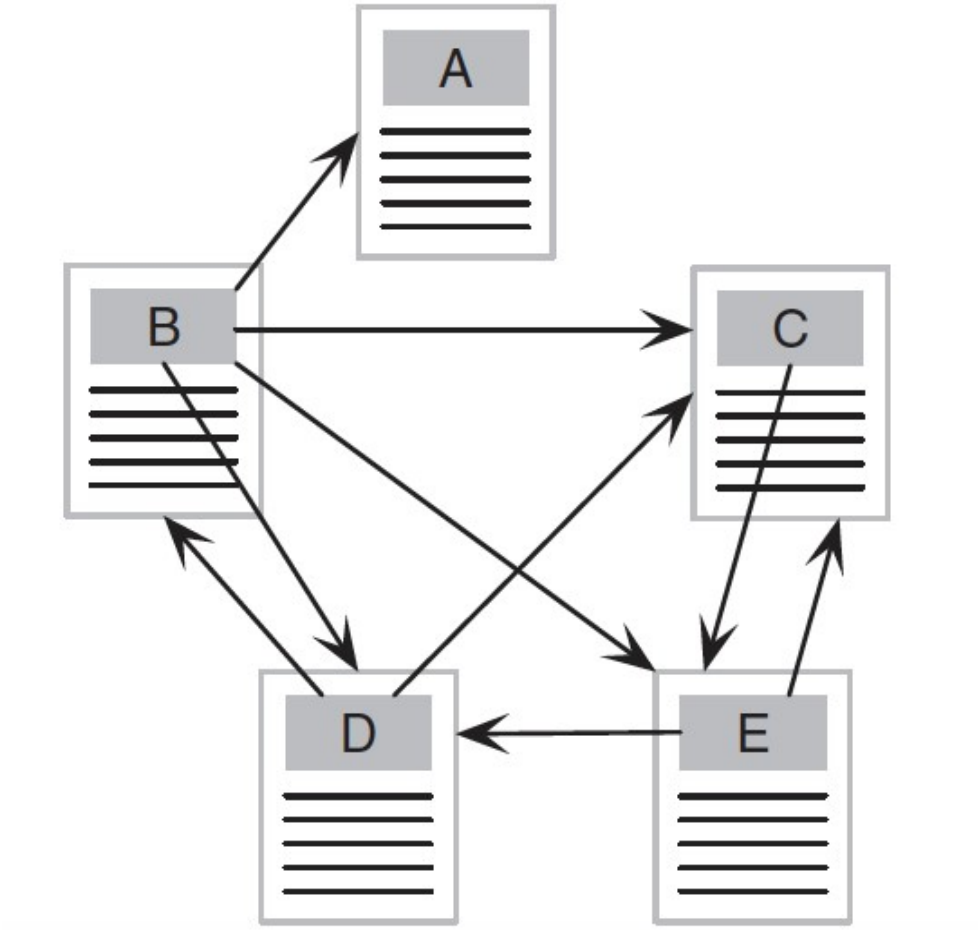

Figure 12: The figure below: Example of PR [1].

Table 3: Table showing limitation of Ranking Models

Ranking Models	Limitations
Boolean Ranking Model	Can't receive partial matches
	The recoup documents are not matched; it cannot predict the degree of relevance
	It does not use term weights
	It can only predict whether a document is relevant to the query terms or not [7,8]
VSM [9].	Doesn't capture semantics
	Cannot denote the "clear logic view" like a Boolean model.
	Avoids the assumption that terms are independent of each other.
LSI [8].	This Assumed that the term is independent of others.
PageRank [15].	It shouldn't be used as a standalone metric; it should be used as parameter only.
	The ranks favours older page, since it is believed that new page including the great ones might not likely have great links to it except is part of already established site.
HITS [14].	Topics drifts and efficient problems occur.
	Non-relevant documents can be recoup.

PageRank with N-Star Ranking Model

Anh, et al. provides a ranking system based on N-linear model with PR to rank SRPs beside authors and, conferences. The system has two models SD4R, and SD3R to rank datasets without citation information and SD4R. The models were tested employing dataset built form DBLP. Sohn, B.S. et al. also provides a generalized network analysis approach to rank SRPs employing N-star model. Based on this model and PageRank algorithm, two different ranking methods were derived, a query/topic independent rank called Universal-Publication rank (UP rank), and a query/topic dependent rank called Topic Publication-rank (TP rank). The model takes into account the mutual relationship between keyword, publication, and tags.

PageRank with HITS Model

Due to M, et al. provides the PageRank algorithm, an extension of PageRank and HITS ranks model. The fundamental knowledge is to measure the relativity between varsities. It measures the indirect relationships between these varsities employing relativity measurements instead of the simple direct citation. Jiang X, et al. provide Mutual-Rank a graph based ranking framework that integrates mutual reinforcement relationships among networks of scholars, researchers, and varsities to achieve a more synthetic, accurate and fair ranking result than previous graph-based methods.

Wang, Y. et al. provide a PageRanks HITS framework that employed various kind of information shared and examined through their usability in any task. HITS is another popular link-based ranking algorithm. Its major difference with PageRank is that a specific type of nodes called Hubs is created and used.
The authors claimed that good authorities are not necessary to point to other good authorities, but good authorities should be linked by many good hubs. A good hub points to many good authorities. In HITS, each node vi has two scores: authority score ai and hub score hi. It can be formally presented as:

$$ai = \sum_{uj \in In(vi)} hj,$$

$$hj = \sum_{vi \in Out\ (vi)} ai.$$

Co-HITs is another link analysis algorithm designed over a bipartite graph with content from two types of objects. The intuition behind the score propagation is the mutual reinforcement to boost co-linked objects. Given a bipartite graph G = (U V, E), where U and V are two disjoint set of vertices.

We use w_{ij}^{uv} (or w_{ji}^{vu}) to denote the weight for the edge between u_i and v_i. To put all the weights between sets U and V together, we can use $w^{uv} \in \mathbb{R}^{[u] \times [v]}$ (or $w^{vu} \in \mathbb{R}^{[v] \times [u]}$) to denote the weight matrix between U and V. For each $u_i \xi$U, a transition probability P_{ij}^{uv} is defined as the probability that vertex u_i in U reaches vertex v_j in V at the next step.

Formally, it is defined as a normalized weight $P_{ij}^{uv} = \frac{w_{ij}^{uv}}{\sum_k w_{jk}^{uv}}$ such that $\sum_{j \varepsilon v} p_{ij}^{uv} = 1$

Similarly, we obtain the transition probability $P_{ji}^{vu} = \frac{w_{ji}^{vu}}{\sum_k w_{jk}^{uv}}$ such that $\sum_{i \varepsilon u} p_{ji}^{vu} = 1$ for each.

Then the iterative framework of Co-HITS can be formulated as:

$r_{(ui)} = (1 - \lambda u) r^o(u_i) + \lambda u \sum_{j \in V} P_{ji}^{vu} r(v_j)$,

$r_{(vj)} = (1 - \lambda v) r^o(v_j) + \lambda v \sum_{i \in U} P_{ij}^{uv} r(u_i)$.

Where, Where λu [0, 1] and λv [0, 1] are personalized parameters, $r^o(u_i)$ and $r^o(v_j)$ are initial ranking scores for u_i and v_j, and $r_{(ui)}$ and $r_{(vj)}$ denote updated ranking scores of vertices u_i and v_j. When both λu and λv are set to be 1, then it becomes the HITs algorithm. And when one of the parameters λu or λv is set to be 1, then it becomes the personalized PageRank.

5. SIMILARITY MEASUREMENT

Similarity measurement is crucial in this thesis since it is directly related to individualized search and morph resolution tasks. Accurate similarity measurement approaches also enable us to construct cleaner networks. We review several commonly used graph-based similarity measures for link prediction. Given a graph G = (V, E), where V is a set of nodes, and E is the set of existing links. Then the following measures can be used to predict the probability of linkage between two nodes x and y. Each of them provides different angles to measure the similarity between two nodes. When there is labeled data available, supervised approaches such as learning to rank algorithms can be leveraged to combine them. Common Neighbors. It measures the size of the common neighbor set between x and y. In other words, sim (x, y) = $|!(x) \backslash !(y)|$, where !(x) and !(x) are neighbor sets for x and y, and |.| is the size of a set. Jaccards coefficient It is a commonly used similarity measures, that can be formulated as: sim (x, y) = $\frac{|!(x) \backslash !(y)|}{|!(x)[!(y)|}$

Adamic/Adar. It aims to capture the importance of each common neighbor. It refines simple counting of common neighbors by putting lower weights on more frequent neighbors, that can be formulated as sim (x,y) =$\sum_{z\in\zeta(x)\eta\zeta(y)} \frac{1}{Log\zeta(z)}$

The above measures are based on neighbor sets, and the following are path-based measures. Path Count. It measures the number of path instances between x and y.Random Walk. It measures the probability of a random walk that starts from x and ends at y. SimRank. SimRank is also a random walk based approach with the assumption that two similar nodes should share many similar neighbors. Normalized Google Distance (NGD). NGD was originally invented to measure the similarity between words and phrases based on their co-occurrence in bulk-scale documents.

The above measures have also been adapted to heterogeneous information networks with the usage of Meta path concept. Let P be a specific type of meta path between two objects x and y, we summarize some of the similarity measures based on P. Different from above measures, meta path-based methods only leverage one type of path each time to compute a specific feature, thus they can capture the unique contributions of each path. Path Count. It measures the number of path instances of P between x and y. specifically, sim(x, y) = $\rho\epsilon P$

5.1 Recall and Precision

In IR systems, precision is often seen as the ability to recover the mst important set of resorurces from available document. Often computed via dividing the exact amount of overlaps between recoup and important sets through numbers of document recouped and as revealed by the equation as follow:

Precision=$\frac{[Relevantset]\cap[Retrievedset]}{[Retrievedset]}$

The recall entails the possibility of providing maximum amount of essential web services from sets of important web-services. It is often calculated through the ratio of the the amount of overlapping between the recoup set of documents and relevant set by the number of relevant sets:

Recall=$\frac{[Relevantset]\cap[Retrievedset]}{[Relevantset]}$

As such, the measures are employed in IR within the basic categorization (For instance, seminal/ non-seminal) that measures the importance of different sets of recouping item thereby evaluating their performance in the information retrieving approach. The amended version of the very measure was presented to examine the web services ranked method. As argued by Bohen, recalling is usually computed through taking the ratio of the highest ranks score by total ranked score of all paper in, as given below.

Precision= $\frac{HighestRankscore}{TotalRankscoreofalluniversities}$

The recalls is achieved by taking the ratio of the highest ranked score and the score of the second ranked algorithms

Recall= $\frac{HighestRankscore}{Scoreof\ 2ndhighestalgorithm}$

6. CONCLUSION

The field of information retrieval and search is very active and attracts a lot of research efforts. Ranking algorithms, being particularly important components of any IR system, affect the efficiency of the information retrieval process. Over the years, work has been done and is still being done in developing better ranking algorithms, as well as optimizing existing ones, to provide a better searching experience for the users. However, there are problems associated with the existing algorithms – among them prohibitive computational complexity, and generalization of search results without reference to user's preferences.

In this paper, a systematic review of the state of technology with regards to information search models and algorithms was outlined. One of the aims of this research work is to develop a computationally simple search model that returns personalized results based on user preferences, i.e. a ranking algorithm that provides solutions to the two problems outlined above. Having given a comprehensive literature review on ranking algorithms, a new ranking algorithm will be discussed in the next chapter, in that the methodology of the provides solution is outlined.

BIBLIOGRAPHY/WORKS CONSULTED/CITED

1. Ankur, G & Rajni, J (2008). An overview of ranking algorithms for search engines. Proceedings of the 2nd National Conference; INDIACom-2008, New Delhi. Feb 08 – 09.
2. Arasu, A; Novak, J; Tomkins, A; Tomlin, J (2002). PageRank computation and the structure of the web: Experiments and algorithms. Proceedings of the Eleventh International World Wide Web Conference, Poster Track. Brisbane, Australia. 107-117.
3. Baeza-Yates, R; Saint-Jean, F & Castillo, C (2002). Web Dynamics, structure and page ranking. Proceedings of 9th International Symposium on String Processing and Information Retrieval, SPIRE, Springer LNCS, Lisbon, Portugal, 117–130.
4. Bates, M, J; Wilde, D, N, & Siegfried, S (1993). An analysis of search terminology used by humanities scholars: The Getty Online Searching Project report no. 1. Library Quarterly, 63(1), 1-39
5. Bates, M. J (1979). Information search tactics. Journal of the American Society for Information Science, 30(4), 205-214.
6. Bates, M; Idea, J. (1979). Tactics. Journal of the American Society for Information Science, 30 (5), 280-289.
7. Beel, J; Gipp, B; Stiller, J (2009). Information retrieval on mind maps – what could it be good for? Proceedings of the 5th International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom'09). Washington, DC: IEEE.
8. Belkin, N, J & Croft, W, B (1987). "Retrieval Techniques," in Annual Review of Information Science and Technology, ed. M. Williams. New York: Elsevier Science Publishers, 109-145.
9. Belkin, N. J (1996). Intelligent information retrieval: whose intelligence? In Proceedings of the 5th International Symposium for Information Science (ISI '96): Humboldt-Universität zu Berlin, 17. -19. Oktober 1996; Krause, J., Herfurth, M., Marx, J., Eds.; Universitätsverlag Konstanz: Konstanz, Germany, 25-31.
10. Belkin, N. J. (1993). Interaction with texts: Information retrieval as information seeking behaviour. In Information Retrieval '93: Von der Modellierung zur Anwendung, Knorz, G., Krause, J., Womser-Hacker, C. Eds.; Universitaetsverlag Konstanz: Konstanz, Germany, 1993; 55-66.
11. Belkin, N. J.; Cool, C.; Stein, A.; Thiel, U (1995). Cases, scripts and information seeking strategies: On the design of interactive information retrieval systems. Expert Systems with Applications, 9 (3), 379-395.

12. Belkin, N. J; Marchetti, P, G; Cool, C (1993). Design of an interface to support user interaction in information retrieval. Information Processing and Management, 29 (3), 325-344.
13. Bellardo, T (1985). What do we really know about online searchers? Online Review, 9 (3), 223-239.
14. Bhavnani, S. K (2002). Important cognitive components of domain-specific search knowledge. In The Tenth Text REtrieval Conference, TREC-2001; Voorhees, E.M.; Harman, D.K. Eds.; Information Today: Medford, NJ, 571-578.
15. Bilal, D (2002). Perspectives on children's navigation of the World Wide Web: Does the type of search task make a difference. Online Information Review, 26 (2), 108-177.
16. Bjørklund, T. A., Götz, M., Gehrke, J., & Grimsmo, N. (2011, October). Workload-aware indexing for keyword search in social networks. In Proceedings of the 20th ACM international conference on Information and knowledge management (pp. 535-544). ACM
17. Bonacich, P (1972). Factoring and weighting approaches to status scores and clique detection. Journal of Mathematical Sociology, pages 113–120
18. Brereton, P., Kitchenham, B. A., Budgen, D., Turner, M., & Khalil, M. (2007). Lessons from applying the systematic literature review process within the software engineering domain. Journal of Systems and Software, 80(4), 571-583.
19. Broder A. Z., Kumar R., Maghoul F., Raghavan P., Rajagopalan S., Stata R., Tomkins A. & Wiener J.L. (2000). "Graph structure in the Web", Proceedings of the 9th International World Wide Web Conference, Amsterdam, The Netherlands, pp. 309-320.
20. Bruza, P. D; Dennis, S (1997). Query-reformulation on the Internet: Empirical data and the hyperindex search engine. In RIAO 97: Conference proceedings with prototype and operational systems demonstrations: Computer-assisted information searching on Internet, McGill University, Montreal, Quebec, Canada, 25th-27th June 1997; RIAO 97, Ed.; CID: Paris, 1997; Vol. 1, 488-499.
21. Bryman, A. (2012). Social research methods. OUP Oxford.
22. Byström, K (2002). Information and information sources in tasks of varying complexity. Journal of the American Society for Information Science and Technology, 53 (7), 581-591.
23. Byström, K; Järvelin, K (1995). Task complexity affects information-seeking and use. Information Processing and Management, 31(2), 191-213.
24. Callahan, E. (2005). Interface design and culture. Annual Review of Information Science and Technology, 39, 257-310.
25. Chang, S (1995). Toward a multidimensional framework for understanding browsing. Unpublished doctoral dissertation, Rutgers University: New Brunswick, N, J.
26. Chen, H & Dhar, V (1991). Cognitive processes as a basis for intelligent retrieval system design. Information Processing and Management, 27 (5), 405-432.
27. Cheng, A & Friedman, E (2006). Manipulability of PageRank under sybil strategies. In First Workshop on the Economics of Networked Systems (NetEcon06), URL http://www.cs.duke.edu/nicl/netecon06/papers/ne06-sybil.pdf.
28. Cheng, Y., Park, J., & Sandhu, R. (2012). A user-to-user relationship-based access control model for online social networks. Data and Applications Security and Privacy XXVI, 8-24.
29. Chu, H (2003). Information Representation and Retrieval in the Digital Age; Information Today: Medford, N. J
30. Cole, C (2001). Intelligent information retrieval: Part IV. Testing the timing of two information retrieval devices in a naturalistic setting. Information Processing and Management, 37 (1), 163-182.
31. Craswell, N., Hawking, D & Robertson, S. (2001). Effective site finding employing link anchor information. In Proceedings of the ACM Conference on Information Retrieval (SIGIR '01), 250-257.
32. Dalal, M. (2007). Personalized social & real-time collaborative search. In Proceedings of the International World Wide Web Conference (WWW '07), 1285-1286.

33. Dieste, O., Grimán, A., Juristo, N., & Saxena, H. (2011, September). Quantitative determination of the relationship between internal validity and bias in software engineering experiments: consequences for systematic literature reviews. In Empirical Software Engineering and Measurement (ESEM), 2011 International Symposium on (pp. 285-294). IEEE.
34. Drabenstott, K. M (2003). Do nondomain experts enlist the strategies of domain experts? Journal of the American Society for Information Science and Technology, 54 (9), 836-854.
35. Duhan, N., Sharma, A. K., & Bhatia, K. K. (2009, March). Page ranking algorithms: a survey. In Advance Computing Conference, 2009. IACC. IEEE International (pp. 1530-1537). IEEE.
36. Dumais, S, T; Belkin, N, J (2005). The TREC interactive tracks: Putting the user into search. In TREC: Experiment and Evaluation in Information Retrieval; Voorhees, E.M.; Harman, D.K., Eds.; The MIT Press: Cambridge, MA, 123-152.
37. Earhart, S. (1986). The UNIX Programming Language, vol. 1. New York: Holt, Rinehart, and Winston.
38. Ellis, D & Haugan, M (1997). Modeling the information seeking patterns of engineers and research scientists in an industrial environment. Journal of Documentation, 53 (4), 384-403.
39. Ellis, D. A (1989). Behavioural approach to information retrieval system design. Journal of Documentation, 45 (3), 171-212.
40. Faloutsos, C. (1985). "Access Methods for Text," Computing Surveys, 17(1), 49-74.
41. Fenichel, C. H (1981). Online searching: Measures that discriminate among users with different types of experience. Journal of the American Society for Information Science, 32 (1), 23-32.
42. Fidel, R & Pejtersen, A, M (2004). From information behavior research to the design of information systems: The cognitive work analysis framework. Information Research, 10 (1). http://informationr.net/ir/10-1/paper210.html.
43. Fidel, R & Soergel, D (1983). Factors affecting online bibliographic retrieval: A conceptual framework for research. Journal of the American Society for Information Science, 34 (3), 163-180.
44. Fidel, R (1985). Moves in online searching. Online Review, 9 (1), 61-74
45. Foote, Jonathan (1999). "An overview of audio information retrieval". Multimedia Systems. Springer.
46. Ford, N, Wilson, T, Foster, D, Ellis, A, & Spink, A, D (2002). Information seeking and mediated searching. Part 4.Cognitive styles in information seeking. Journal of the American Society for Information Science and Technology, 53 (9), 728-735.
47. Ford, N; Miller, D, & Moss, N (2002). Web search strategies and retrieval effectiveness: An empirical study. Journal of Documentation, 58 (1), 30-48.
48. Frakes, W, B. (1992). Information retrieval data structures & algorithms. Prentice-Hall, Inc
49. Franceschet, M (2002). "PageRank: Standing on the shoulders of giants".
50. Fujimura K., Inoue R., and Sugisaki M. (2005). "EigenRumor Algorithm for Ranking Blogs". May 10-14, 2005, Chiba, Japan.
51. G. Jeh, G & Widom, J (2003). Scaling personalized web search. In Proceedings of the International World Wide Web Conference ('03), 271-279, 2003.
52. García-Crespo, Á, Colomo-Palacios, R., Gómez-Berbís, J. M., & García-Sánchez, F. (2010). SOLAR: social link advanced recommendation system. Future Generation Computer Systems, 26(3), 374-380
53. Ghafari, M., Saleh, M., & Ebrahimi, T. (2012). A federated search approach to facilitate systematic literature review in software engineering. International Journal of Engineering, 23(5), 12-36
54. Goodrum, Abby A. (2000). "Image Information Retrieval: An Overview of Current Research". Informing Science. 3 (2).
55. Google technology overview (2004) "http://www.google.com/intl/en/corporate/tech.html.

56. Google. Personalized search for everyone. http://googleblog.blogspot.com/2009/12/personalized-search-for-everyone.html. Last recoup on April, 16, 2010.
57. Greenwald, A., & Wicks, J. (2006). QuickRank: A recursive ranking algorithm. In Proc. of the 1st International Workshop on Computational Social Choice.
58. Greenwald, A., & Wicks, J. (2006). QuickRank: A recursive ranking algorithm. In Proc. of the 1st International Workshop on Computational Social Choice.
59. Gupte, M., Shankar, P., Li, J., Muthukrishnan, S., & Iftode, L. (2011, March). Finding hierarchy in directed online social networks. In Proceedings of the 20th international conference on World wide web (pp. 557-566). ACM.
60. Gyongyi Z. & Garcia-Molina H. Web spam taxonomy (2004). Technical report, Stanford University Technical Report.
61. Haveliwala, T. (2002). "Topic-Sensitive PageRank", In Proceedings of the 11th World wide Web Conference.
62. Hawk, W. B.; Wang, P (1999). Users' interaction with the World Wide Web; Problems and problem solving. Proceedings of the 62nd ASIS Annual Meeting, 36, 256-270.
63. Hema, D, B. & Roy, N. (2011). An improved Page Rank Algorithm based on Optimized Normalization Technique. International Journal of Computer Science and Information Technologies, 2 (5), 2183-2188
64. Hotho, A; Jäschke, R; Schmitz, C, & Stumme, G. (2006). Information retrieval in folksonomies: Search and ranking. In Lecture Notes in Computer Science. Vol. 4011/2006(the Semantic Web: Research and Applications), 411-426
65. Howard, H (1982). Measures that discriminate among online users with different training and experience. Online Review, 6 (4), 315-326.
66. Hsieh-Yee, I (1993). Effects of search experience and subject knowledge on the search tactics of novice and experienced searchers. Journal of the American Society for Information Science, 44 (3), 161-174.
67. Hyldegard, J (2006). Collaborative information behaviour: Exploring Kuhlthau's information search process model in a group-based educational setting. Information Processing and Management, 42 (1), 276-298.
68. Ingwersen, P & Järvelin, K (2005). The Turn: Integration of Information Seeking and Retrieval in Context; Springer. Germany: Heidelberg
69. Ingwersen, P (1992). Information retrieval interaction; Taylor Graham: London.
70. Ingwersen, P (1996). Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. Journal of Documentation, 52 (1), 3-50.
71. J.-T. Sun, H.-J. Zeng, H. Liu, Y. Lu & Z. Chen. (2005). Cubesvd: A novel approach to personalized web search. In Proceedings of the International World Wide Web Conference (WWW'05), pp. 382-390.
72. Jamshidi, P., Ghafari, M., Aakash, A., & Pahl, C. (2012). A protocol for systematic literature review on Architecture-Centric Software Evolution Research. Technical Report, Lero-The Irish Software Engineering Research Centre, Dublin City University.
73. Jansen, B. J. & Rieh, S. (2010). The Seventeen Theoretical Constructs of Information Searching and Information Retrieval. Journal of the American Society for Information Sciences and Technology. 61(8), 1517-1534.
74. Jansen, B. J.; Spink, A.; Saracevic, T (2000). Real life, real users, and real needs: A study and analysis of user queries on the Web. Information Processing and Management 2000, 36 (2), 207-227.
75. Joachims, T. (2002). Optimizing search engines employing clickthrough data. In Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (SIGKDD'02), 133-142

76. Jones, K. S., Walker, S & Robertson, S. E (2000). A probabilistic model of information retrieval: Development and comparative experiments. Information Processing and Management. 36 (6), 779-808, 2000.
77. Jones, S.; Cunningham, S.J.; McNab, R.; Boddie, S (2000). Human-computer interaction for digital libraries: A transaction log analysis of a digital library. International Journal on Digital Libraries , 3 (2), 152-169.
78. K. Bharat and G.A. Mihaila (2002). When experts agree: Employing Non-Affiliated Experts to Rank Popular Topics. ACM Transactions on Information Systems, 20(1), 47-58,.
79. Kitchenham, B., Pretorius, R., Budgen, D., Pearl Brereton, O., Turner, M., Niazi, M., & Linkman, S. (2010). Systematic literature reviews in software engineering–a tertiary study. Information and Software Technology, 52(8), 792-805.
80. Kleinberg J. M. (1999). Authoritative sources in a hyperlinked environment. Journal of the ACM, 46(5), 604 –632.
81. Kracker, J (2002). Research anxiety and students' perceptions of research: An experiment: Part 1. Effect of teaching Kuhlthau's ISP model. Journal of the American Society for Information Science and Technology, 53 (4), 282-294.
82. Kuhlthau, C. C. (1991). Inside the search process: Information seeking from the user's perspective. Journal of the American Society for Information Science 1991, 42 (5), 361-371.
83. Kumar, R., Novak, J., & Tomkins, A. (2010). Structure and evolution of online social networks. Link Mining: Models, Algorithms, and Applications, 337-357.
84. L. Srour, A Kayssi and A. Chebab (2007). Personalized Web Page Ranking Employing Trust and Similarity", IEEE 2007.
85. Lau, T.; Horvitz, E (1999). Patterns of search: Analyzing and modeling Web query refinement. In Proceedings of the 7th International Conference on User Modeling Banff, Canada, June 1999; Kay, J., Ed.; Springer-Wien: New York, 119-128.
86. Lazonder, A. W.; Biemans, H. J. A.; Wopereis, I. G. J. H (2000). Differences between novice and experienced users in searching information on the World Wide Web. Journal of the American Society for Information Science, 51 (6), 576-581.
87. Lempel R. & Moran S. (2000). The stochastic approach for link-structure analysis (SALSA) and the TKC effect. Proceedings of the 9th. International World Wide Web Conference, Amsterdam, The Netherlands, pp. 387 – 401
88. Liu N. C. and Cheng Y. (2005). The academic ranking of world universities. Higher Education in Europe, 30(2), 1-14
89. M. Montaner, B. López, & J. L. de La Rosa (2003). A taxonomy of recommender agents on the internet. Artificial Intelligence Review. 19(4), 285-330.
90. M. Richardson and P. Domingos (2002). The Intelligent Surfer: Probabilistic Combination of Link and Content Information in PageRank", Advances in Neural Information Processing Systems, 14, 1441-1448, MIT Press.
91. M.S. Aktas, M.A. Nacar, and F. Menczer (2004). Personalizing PageRank Based on Domain Profiles, WebKDD.
92. Manning, C., Raghavan, P & Schutze, H (2008). Introduction to information retrieval. Cambridge: University Press.
93. Marchionini, G (1995). Information-Seeking in electronic environments; Cambridge University Press: Cambridge,.
94. Marchionini, G.; Dwiggins, S.; Katz, A.; Lin, X (1993). Information seeking in full-text end-user-oriented search-systems - The roles of domain and search expertise. Library and Information Science Research, 15 (1), 35-69.
95. Markey, K., Atherton, P (1978). Online training and practice manual for ERIC Database Searchers; ERIC Clearinghouse on Information Resources: Syracuse, NY

96. Marques, A. B., Rodrigues, R., & Conte, T. (2012, August). Systematic Literature Reviews in Distributed Software Development: A Tertiary Study. In Global Software Engineering (ICGSE), 2012 IEEE Seventh International Conference on (pp. 134-143). IEEE.
97. Micarelli, A. Gasparetti, F; Sciarrone, F & Gauch, S (2007). Personalized search on the World Wide Web. In Lecture Notes in Computer Science. Volume 4321 (the Adaptive Web), pp. 195-230, 2007.
98. Mislove, A; Gummadi, K. P & Druschel, P (2006). Exploiting social networks for internet search. In Proceedings of the 5th Workshop on Hot Topics in Networks (HotNets'06).
99. Montenegro R. & Tetali P. (2006). Mathematical aspects of mixing times in Markov chains. Foundations and Trends in Theoretical Computer Science, 1(3), 237 – 354.
100. Moukdad, H.; Bulk, A (2001). Users' perceptions of the Web as revealed by transaction log analysis. Online Information Review, 25 (6), 349-359.
101. Novak, J. D. (1998). Learning, creating, and employing knowledge: Concept maps as facilitative tools in schools and corporations. Mahwah, NJ: Lawrence Erlbaum Associates.
102. Novak, J. D., & Gowin, D. B. (1984). Learning how to learn. New York, NY: Cambridge University Press.
103. Novak, J., Fleishmann, M., Strauss, W., Schneider, M., Wurst, M., Morik, K., & Kunz, C. (2002). Augmenting the knowledge bandwidth and connecting heterogeneous expert.
104. Okoli, C. and Schabram, K. (2010). A Guide to Conducting a Systematic Literature Review of Information Systems Research. Sprouts. Working Papers on Information Systems, 10(26).
105. P. A. Chirita, W. Nejdl, R. Paiu & C. Kohlschütter (2001). Employing odp metadata to personalize search. In Proceedings of the ACM Conference on Information Retrieval (SIGIR '01), pp. 250-257.
106. Palmquist, R. A., & Kim, K. S (2000). Cognitive style and online search experience on Web search performance. Journal of the American Society for Information Science and Technology, 51 (6), 558-567.
107. Pennanen, M., Vakkari, M, & Students', P. (2003). Conceptual structure, search process and outcome while preparing a research proposal. Journal of the American Society for Information Science, 54 (8), 759-770.
108. Prasad Chebolu & Pall Melsted (2008). "PageRank and the random surfer model", Proceedings of 19th annual ACM-SIAM Symposium on Discrete Algorithms, pp. 1010-1018
109. Ready, R. C & Hu, D. (1995). Statistical Approaches to the Fat Tail Problem for Dichotomous Choice".
110. Rieh, S. Y.; Xie, H (2006). Analysis of multiple query reformulations on the web: The interactive information retrieval context. Information Processing & Management, 42 (3), 751-768.
111. Rosvall, M., & Bergstrom, C. T. (2010). Mapping change in bulk networks. PloS one, 5(1), e8694.
112. Rosvall, M., & Bergstrom, C. T. (2011). Multilevel compression of random walks on networks reveals hierarchical organization in bulk integrated systems. PloS one, 6(4), e18209.
113. S. Brin & L. Page (1998). The anatomy of a bulk-scale hypertextual Web search engine, In Proceedings of the International World Wide Web Conference (WWW '98), pp. 107-117, 1998. International Journal on Web Service Computing (IJWSC), 3(1),
114. Salton, G & Buckley, C (1988). Term-weighting approaches in automatic text retrieval. Information Processing and Management. 24(5), 513-523.
115. Saracevic, T (1996). Modeling interaction in information retrieval (IR): A review and proposal. Proceedings of the 59th ASIS Annual Meeting 1996, 33, 3-9.
116. Saracevic, T (1997). The stratified model of information retrieval interaction: Extension and applications. Proceedings of the 60th ASIS Annual Meeting 1997, 34, 313-327.

117. Schacter, J.; Chung, G. K. W. K.; Dorr, A (1998). Children's Internet searching on complex problems: Performance and process analyses. Journal of the American Society for Information Science, 49 (9), 840-849.
118. Schenkel, T. Crecelius, M. Kacimi, S. Michel, T. Neumann, J. X. Parreira, X. & Weikum, G. (2008). Efficient top-k querying over social-tagging networks. In Proceedings of the ACM Conference on Information Retrieval (SIGIR '08), pp. 523-530, 2008.
119. Shiri, A. A.; Revie, C (2003). The effects of topic complexity and familiarity on cognitive and physical moves in a thesaurus-enhanced search environment. Journal of Information Science, 29 (6), 517-526.
120. Shute, S. J.; Smith, P. J (1993). Knowledge-based search tactics. Information Processing and Management, 29 (1), 29-45.
121. Siegfried, S.; Bates, M.J.; Wilde, D.M (1993). A profile of end-user searching behavior by humanities scholars: The Getty online searching project (Rep. No. 2). Journal of the American Society for Information Science, 44 (5), 273-291.
122. Silverstein, C.; Henzinger, M.; Marais, H.; Morica, M (1999). Analysis of a very bulk Web search engine query log. SIGIR Forum, 33 (1), 6-12.
123. Soloman, P (1993). Children's information retrieval behavior: A case analysis of an OPAC. Journal of the American Society for Information Science, 44 (5), 245-264.
124. Spink, A.; Jansen, B. J (2004). Web Search: Public Searching of the Web; Kluwer Academic Publishers: Boston.
125. Spink, A.; Wolfram, D.; Jansen, B. J.; Saracevic, T (2001). Searching the Web: The public and their queries. Journal of the American Society for Information Science, 52 (3), 226-234.
126. Suri, P. & Taneja, H. (March 2012). An Integrated Ranking Algorithm for Efficient Information Computing in Social Networks. International Journal on Web Service Computing (IJWSC), 3(1), 31 – 44.
127. Sutcliffe, A. G.; Ennis, M.; Watkinson, S. J (2000). Empirical studies of end-user information searching. Journal of the American Society for Information Science, 51 (13), 1211-1231.
128. Vakkari, P (2000). ECognition and changes of search terms and tactics during task performance: A longitudinal study. In RIAO' 2000 Conference Proceedings, Content-Based Multimedia Information, Collège de France, Paris, France, April 12-14, 2000; RIAO, Eds.; C.I.D.: Paris, 2000; 1, 894-907. http://www.info.uta.fi/vakkari/Vakkari_Tactics_RIAO2000.html (accessed July 15, 2018).
129. Vakkari, P (2000). Relevance and contributory information types of searched documents in task performance. In Proceedings of the 23rdAnnual International ACM SIGIR Conference on Research and Development in Information Retrieval; Belkin, N.J., Ingwersen, P., Leong, M-K, Eds; SIGIR forum; ACM Press: New York, Vol. 34, 2-9.
130. Vakkari, P. A (2001). Theory of the task-based information retrieval process. Journal of Documentation, 57 (1), 44-60.
131. Vakkari, P.; Hakala, N (2000). Changes in relevance criteria and problem stages in task performance. Journal of Documentation, 56, 540-562.
132. Vakkari, P.; Pennanen, M.; Serola, S (2003). Changes of search terms and tactics while writing a research proposal. Information Processing and Management, 39 (3), 445-463.
133. W. B. Frakes & R. Baeza-Yates. Information Retrieval: Data Structures & Algorithms, (2009).
134. Walker, G.; Janes, J (1999). Online Retrieval: A Dialogue of Theory and Practice, 2nd Ed.; Libraries Unlimited: Englewood, Colorado.
135. Wang, P.; Berry, M.; Yang, Y (2003). Mining longitudinal Web queries: Trends and patterns. Journal of the American Society for Information Science and Technology, 54 (8), 743-758.
136. Wang, P.; Hawk, W. B.; Tenopir, C (2000). Users' interaction with World Wide Web resources: An exploratory study employing a holistic approach. Information Processing & Management, 36 (2), 229-251.

137. Webometrics Website (2016).
138. Wildemuth, B. M (2004). The effect of domain knowledge on search tactic formulation. Journal of the American Society for Information Science and Technology, 55 (3), 246-258.
139. Wilson, T. D (2000). Human information behaviour. Informing Science, 3(2), 49-56.
140. Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., & Wesslén, A. (2012). Systematic Literature Reviews. Experimentation in Software Engineering, 45-54.
141. Wolfram, D.; Xie, H (2002). Traditional IR for Web users: A context for general audience digital libraries. Information Processing & Management, 38 (5), 627-648.
142. X. Shen., B.Tan, & C. Zhai, C (2002). Ucair: Capturing and exploiting context for personalized search. In Proceedings of the Information Retrieval in Context Workshop, SIGIR IRiX'05, 2005.
143. Xie, H (2000). Shifts of interactive intentions and information-seeking strategies in interactive information retrieval. Journal of the American Society for Information Science, 51 (9), 841-857.
144. Xie, H (2002). Patterns between interactive intentions and information-seeking strategies. Information Processing & Management, 38 (1), 55-77.
145. Xie, H. Understanding human-work domain interaction: Implications for the design of a corporate digital library. Journal of the American Society for Information Science and Technology, 57 (1), 128-143.
146. Xie, I. Interactive information retrieval in digital environments; IGI Global Inc.: Hershey, Pennsylvania, 2008.
147. Xing, W & Ghorbani, A. (2004). "Weighted PageRank algorithm". Proceedings of the Second Annual Conference on Communication Networks and Services Research 21-24 May 2004. Fredericton, Canada.
148. Y. Hu, G. Xin, R. Song, G. Hu, S. Shi, Y. Cao, & H. Li, H. (2005). Title extraction from bodies of html documents and its application to web page retrieval. In Proceedings of the ACM Conference on Information Retrieval (SIGIR '05), pp. 250-257.
149. Zareh Bidoki A. and Yazdani N. (2007). "DistanceRank: An intelligent ranking algorithm for web pages". Information Processing and Management (2007), doi:10.1016/j.ipm.2007.06.004