Formal Definitions of Web Information Search

Su Yan
Information Sciences and Technology
The Pennsylvania State University
University Park, PA, 16802
syan@ist.psu.edu

C. Lee Giles
Information Sciences and Technology
The Pennsylvania State University
University Park, PA, 16802
giles@ist.psu.edu

Bernard J. Jansen
Information Sciences and Technology
The Pennsylvania State University
University Park, PA, 16802
jjansen@ist.psu.edu

Su Yan is the corresponding author

Abstract
Research in Web search engines has been criticized for lacking underlying theories and models. Theories adopted from traditional information retrieval research have been found in many ways lacking and inefficient in dealing with information retrieval in the Web context, primarily because of the amount of information and its dynamic nature, the hyperlinked structure, and multimedia sources. Appropriate Web models and theories for search engines will make web search and information retrieval problems easier to formulate and comprehend. This in turn helps to highlighting holes in current Web search engine techniques. We analyze and categorize previous Web and information retrieval models. Grounded on previous work, we then propose a new Web information retrieval model based on both objective and subjective criteria. The performance of the new model is systematically compared with other IR models, and contributions of this work are highlighted.

Introduction

Since its advent, the World Wide Web (WWW) has become one of the largest and most readily accessible repositories of human knowledge. Its scholarly, scientific, and business applications have been staggering. Given the large volume of Web pages, it is no surprise that Internet users are increasingly using search engines and search services to find specific information. The number of web searches in May 2005 alone amounted to 4.3 billion (Nielsen Netratings for search engines 2005). Such use of Web search engines has motivated a large effort to make Web search engines more efficient tools that manage, retrieve, and filter information.

There has been much concern with the lack of underlying theories and models for Web search engines and Web search. Arasu et al. (2001) report that many of the search engines use well-known traditional information retrieval (IR) algorithms and techniques, which were originally developed for relatively small and structured collections such as book catalogs in a (physical) library. However, the Web is a massive, semi-structured (if not unstructured) constantly changing dataset that is high in redundancy (Baeza-Yates & Ribeiro-Neto, 1999). Brooks (2003) reports in a survey of Web technology that the classic IR strategy - indexing with topical metadata, has experienced disappointing results when applied to Web searching. Bianchini et al. (2005) state that most commonly used scoring algorithms for Web searching are directly derived from traditional IR methods. Unfortunately, these methods have limitations and drawbacks since they do not take into account the graphical structure of the Web. Tsikrika et al. (2002) point out similar

problems. They state that "Web IR" adopts models, algorithms, and heuristics previously developed in the traditional hypertext/hypermedia IR environments, which have proved ineffective when applied to the Web searching.

We found that little work has been done regarding the fundamental properties and characteristics of Web search engines. New theories and models are needed to systematically generalize and formalize various Web search algorithms and techniques. We believe that Web search engine formalism is the first step toward building a strong theoretical foundation for the research in Web search. Such formalism should answer basic questions like: "What is a Web search engine?" "What are the basic components of a Web search engine?", and "In what way are the basic components related together?" As noted by others (Goncalves et al. 2004, Tague et al. 1991), formal models are crucial to describe and understand clearly and unambiguously the general characteristics of the complex information systems, explain their structures and processes, and strengthen their common practice in design. In this paper, we propose a new formal model that will contribute to the theoretical and practical unification of Web search engines.

The paper is organized as follows. We first analyze existing classic IR models, clarify concepts regarding IR formalisms, and propose a new categorization for existing IR models. The strengths and shortcomings for each classic IR model are discussed. We then discuss the new challenges faced by Web search engines and propose both objective criteria and subjective criteria for Web search problems. Then we propose a new formal model for Web information search and compare the new model with the existing ones. In case study, PageRank and HITS searching strategies are analyzed to test the explanation power of the new model. Finally, we conclude the paper with a discussion about the limitation of current work and insight into future work.

Categorization of formal IR models

A Web search engine is an IR system in a Web context (Meng et al. 2002). IR provides a theoretical basis as well as practical examples for the Web search engine study. However, ambiguities exist in IR formalisms. Model or modeling is mentioned in almost every paper within IR research. Different people may indicate quite different concepts by using the same term model. The confusion caused by the popular usage of the term model in IR area has also been studied by other researchers. Crestani and Lalmas (2001) differentiate model from meta-model. Baeza-Yates and Ribeiro-Neto (1999) propose the concept of "formal characterization of IR models" which refers to the same concept as meta model in (Crestani & Lalmas, 2001). Grossman and Frieder (2004) differentiates IR model from retrieval strategy, etc. However, none of the previous work considers all the different concepts of the term "model" in IR work. This causes even more confusions. To compare and study previous work precisely, we generalize existing IR models into three major categories: *definition-oriented (DO) IR models, strategy-oriented (SO) IR models, explanation-oriented (EO) IR models.* This paper focuses on DO models.

Definition-oriented (DO) models

A DO model formally defines the entities, relationships, and operations, which form the systems that the model intends to describe. A complete DO model will contain a representation of all components for any system of the kind referenced by the model.

Grossman-Frieder (GF) model

The GE model (Grossman & Frieder 2004) defines an IR system as a triple (original notations are used):

$I = (D, R, \delta)$, where

D is a document collection;

R is the set of queries;

δ is the retrieval function. $\delta_j : R_j \to 2^{D_j}$, $\delta_j \in \delta$

The GF model defines the retrieval process as identifying a subset of documents (2^{D_j}) relevant to a given query. The retrieval function determines which document is relevant. However, no ranking strategy is defined, which leads to the difficulty in explaining ranking-based retrieval strategies with GF model.

User-centered (UC) model

UC model (Dominich 2001) defines an IR system as (original notations are used):

$IR = m\left[\Re\left(O,\left(Q,\left\langle I,\mapsto\right\rangle\right)\right)\right]$, where

O is the set of objects to be retrieved;

Q is the set of queries;

I is the information about the user known in advance;

$\mapsto$ is the information derivable (deductible) from the user information I;

$\Re$ is a relationship between the objects in O and the information need IN;

m represents that the relation $\Re$ is established with some uncertainty.

Accordingly, the user's information need (or user model) is $IN = \left(Q,\left\langle I,\mapsto\right\rangle\right)$.

Although the definition of UC model is not mathematically strict (e.g. the use of "m" and "$\Re$"), it introduces an important concept - the user model, into an IR system. However, the model fails to tell how an IR system should use the user model. Therefore, the UC model is more like an independent IR user model rather than a comprehensive definition for the whole IR system.

Similarity-thesauri-enabled (STE) model

The STE model (Sheridan et al., 1997) enables the construction of the similarity thesauri for cross-language information retrieval. Since thesauri are not considered in this paper, we concentrate on the IR model part of the STE model. According to the STE model, IR is a tuple (original notations are used):

$\left\langle T,\Phi,D;ff,df\right\rangle$, where

T is the set of all tokens (terms) in a document, $\tau \in T$;

Φ is the set of indexing terms/vocabulary, $\phi_i \in \Phi$, $\phi : T \to \Phi, \tau \mapsto \phi(\tau) := \phi_i$, ϕ function maps the set of all tokens to the indexing vocabulary Φ;

D is the set of documents, $d_j \in D$, $d : T \to D, \tau \mapsto d(\tau) := d_j$, d function maps T to the document collection D;

ff is the frequency of an indexing term in a document, $ff(\phi_i, d_j) := \left|\{\tau \in T \mid \phi(\tau) = \phi_i \wedge d(\tau) = d_j\}\right|$;

df is the document frequency of an indexing term, $df(\phi_i) := \left|\{d_j \in D \mid \exists \tau \in T : \phi(\tau) = \phi_i \wedge d(\tau) = d_j\}\right|$.

Despite the detailed definition for the elements of an IR system, the STE model does not define the relationship between elements. For example, there is no operation that actually maps queries to documents, and no function that actually uses ff and df. Due to the nonexistence of that relationship, this model fails to be a strong mathematical framework for IR systems.

Baeza-Yates-Ribeiro-Neto (BYRN) model

According to the BYRN model (Baeza-Yates & Ribeiro-Neto, 1999), an IR system is a quadruple (original notations are used):

$\left\langle D,Q,F,R(q_i,d_j)\right\rangle$, where

D is a set composed of logical views (or representations) for the documents in the collection;

Q is a set composed of logical views (or representations) for the user's information needs. Such representations are called queries;

F is a framework for modeling document representations, queries, and their relationships;

$R(q_i,d_j)$ is a ranking function which associates a real number with a query $q_i \in Q$ and a document representation $d_j \in D$. Such ranking defines an ordering among the documents with regard to the query q_i.

The component F and the ranking function R make the BYRN model flexible to express many existing retrieval strategies. For example, if the framework is a vector space, the BYRN model yields a vector space IR system, where documents and queries are formalized as vectors. Ranking function R makes this model more expressive since most current IR techniques are based on certain ranking strategies. However, the BYRN model is too general to express and solve specific problems.

Strategy-oriented (SO) Models

SO models formally define various retrieval strategies. A retrieval strategy is a specific implementation of a DO model. For example, if a SO model adopts the definition of an IR system given by the DO model BYRN, the SO model will assume that an IR system should contain all the elements and relationships defined by the RYRN model, which are documents, queries, a mathematical/logical framework, and a ranking function. Given these basic elements and relationships, the SO model represents the implementation of the whole retrieval process in a formal way. The three classic IR models (the Boolean model, the vector space model, and the probability model), and their extensions, are examples of SO models.

Explanation-oriented (EO) Models

The EO models are defined to generalize existing SO models, and extract the common procedure and general features among different retrieval strategies in order to strengthen common practice. Most SO models are restricted within a single mathematical framework. The retrieval strategy defined in one SO model is hard to explain, implement, or extend to other SO models. For example, the retrieval strategies defined in the probability IR models are hard to explain in the vector space models. The EO model commonly defines one unifying general mathematical framework, within which it is possible to discuss different retrieval strategies. (van Rijsbergen 2004, van Rijsbergen 2005) proposes the idea of applying Hilbert space mathematics, the language for quantum mechanics, to IR. It is shown that the three classic IR models and possibly others can be described and represented within a unifying framework. For easy reference, we call this model a geometrical model. (Rolleke et al., 2003) propose a general matrix framework (GMF) model for IR. A topological model of IR is introduced in (Everett & Cater 1992, Egghe & Rousseau 1998).

New Challenges

Similar to traditional IR systems, Web search engines deal with three major objects: the information repository, the individual documents, and the user. However, unique attributes of the Web, Web pages, and Web users distinguish Web information search from traditional IR problems.

The Web is a huge and rapidly growing universal knowledge repository. Web search engines face the new *Abundance Problem*: “The number of pages that could reasonably be returned as relevant (by using traditional IR methods) is far too large for a human user to digest” (Kleinberg 1999). The interlinked nature of the Web also sets it apart from other static text corpora, and provides a new perspective to evaluate Web pages. A hyperlink pointing to a Web page represents a particular author's evaluation of that Web page. Thus, the graph structure of the Web encodes the evaluations of any Web pages that are made by a collection of large numbers of independent individuals. We can consider that the Web encodes some objective evaluations of individual Web pages. Accordingly, the “relevance-based” traditional IR retrieval criteria are subjective evaluations.

In addition to the wildly known linkage feature, Web pages are highly tagged. The structural information of HTML and XML documents is easily available through tags. Traditional IR systems typically disregard the structure information of a document, because commonly this information is either not available, or is hard to acquire (Cutler et al., 1997). Multimedia is proliferating on the Web. Frustration with traditional IR in dealing with multimedia retrieval is growing. For example, there is the subjectivity issue in extracting descriptive keywords from images (Jain 1995). The exclusive use of extracted keywords may prove insufficient, especially when the user is interested in the visual components of an image (Kherfi et al., 2004).

Understanding the user and the user's information need is a new challenge posed to Web search. Web search engines are not designed for information experts, as traditional IR systems do, but are instead designed for heterogeneous user group. Experienced and novice users construct searches differently. Web users differ from the user model assumed in traditional IR work (Silverstein et al., 1999). How to make the Web search engine useful for all kinds of users is a major issue.

Formal model for Web search engine (SE model)

As shown in the previous section, adapting traditional IR models to Web search left much to be desired. However, no formal model for Web search engines has been proposed. In this section, we propose a new DO formal model for Web search engines. The new model incorporates both objective and subjective criteria, considers the link structure of the Web, takes special account of the Web user, is media independent, and is capable of taking advantage of the unique structural information of Web pages. Basic concepts and components of Web search engines are defined and explained first.

It needs to be noted that there are other issues differentiate Web search engines from traditional IR systems. For example, there is the issue of dynamics. Both the user's information need and the information repository (Web) change over time, which leads to the dynamics of Web search. There is also the issue of validity. There are users actively on the Web trying to subvert the ranking process. Never has this happened in any other information environment. For our current state of study, we do not consider all these issues. However, we think it is important to include them in future work on Web search engine formalism in order to improve the explanatory power of the model.

Definition 1. An *object* is a pair $OB = \langle l, A_0 \rangle$ where

$l \in L$, L is a set of universally unique labels;

$A_o \in 2^A$, A is the *attribute space* (refer to definition 2).

An *object* is an entity of the real physical world (e.g. Web). It can be any piece of information (e.g., Web pages, images, video or audio clips etc.) that a user may want the search engine to retrieve. An object has its own identity (*label*) that does not change throughout its lifetime. The attribute distinguishes one object from all the others, and enables the unique identity l to be assigned to the object. Normally, individual object cannot possess all the attributes defined by A. For example, a pure textual document cannot have visual properties such as color, shape, texture, etc. Therefore, the attributes that an individual object possesses A_o is a subset of the whole attribute space. As it will be shown later, this definition of object makes the discussion about multimedia objects easier and more flexible.

Definition 2. *Attribute space* is a set containing all the possible physical properties of objects, denoted as. Normally, an attribute space $A = \{A_t, A_m, A_s\}$ is composed of three major subspaces:

- textual attributes A_t: Refer to the properties conveyed by the content of the textual part of an object. Term frequency, inverse document frequency, and document models all are examples of the use of textual attributes. Note that, multimedia objects may also contain textual attributes. For example, keywords extracted from the surrounding text, the page title, the file name etc;
- multimedia content attributes A_m: Refer to the content information conveyed by the multimedia object itself. For example, color, orientation, edges, textures, shapes, etc are commonly used pictorial attributes of visual objects (Kherfi et al., 2004);
- structural attributes A_s: Refer to the structure of the data being considered. For HTML files, the structure information can be extracted from tags. For multimedia objects, the length and width of a piece of image, the number of frames of a video clip are some examples.

Definition 3. A *feature* is a piece of information used to logically/mathematically describe and represent the attributes of the object under consideration in a space. Therefore: Let $F = \{f_1, f_2, \cdots, f_M\}$, $M \geq 1$, be a set for basic features, a feature is defined as $F_o \subset 2^F$. The set of features F together with the operations Op defined on the set make up a space $S = (F, \mathrm{Op})$, which is called a feature space.

"Attributes" and "features" refer to an object's properties in the physical space the abstract logical or mathematical space respectively. For example, all the terms contained in a piece of textual document are the textual attributes, which help readers to understand the content of the document. However, features are only the index terms, which are acquired through case folding, stemming, and stopping. Therefore, features can be real words, or just the roots of words. Similar things happen in image retrieval, where the color features chosen depend on the quantization strategy being adopted (Faloutsos et al., 1994). Because a system can only have a limited understanding capability (Fuhr 1992), features cannot completely express the entire original attributes of objects in the physical space. One key problem in both textual and multimedia retrieval is to find "descriptive" features that can capture more attributes and will not cause a serious curse-of–dimensionality.

Definition 4. Let $\mathrm{I}: 2^{A} \to \left(2^{F}, \mathrm{Op}\right)$ be a *representing function* where F is the set of (indexing) features and A is the set of attributes.
A representing function maps attributes to feature spaces. Because the operations Op are defined on the set of features, we can define measures in the feature space and measure each feature and their relationship. The measurable nature of the feature space is the reason why we use representing functions to map attributes to feature spaces. An attribute can be represented in different ways, depending on the various features being used, and the different operations defined on the set of features. For example, both the color-histogram and edge-orientation histogram are visual features used to represent visual content attributes of images.

Definition 5. Let D be the set of logical/mathematical representation for the objects under consideration in the abstract space (feature space), collectively called *documents*. A document is obtained by applying one or more representing functions to the corresponding object. The process can be expressed as:
$\mathrm{I}: OB \to D$, or
$\mathrm{I}: \langle l, A_o \rangle \to \langle l, (F_d, \mathrm{Op}) \rangle$, where $F_d \subset 2^{F}$.
Documents are objects of interests (to be retrieved) in the logical/mathematical space. Each document has its own properties represented as feature F_d that enable a unique label l assigned to that document. The subscripts o and d describe the space under consideration, either the physical object space, or the abstract document space.

Definition 6. *Indexing* is defined as a collection of representing functions, which are expressed as: $\mathrm{Id} = \{\mathrm{I}_1, \mathrm{I}_2, \cdots, \mathrm{I}_N\}$, $N \geq 1$.
Indexing is a process of deciding the attributes of the physical objects to be represented, deciding the features to be used to do the representation, and doing the representation by mapping physical world attributes to their logical representations in a measurable space. Each representing function $\mathrm{I}_i \in \mathrm{Id}$, $i = [1, N]$ is one mapping method.

For each Web search engine, multiple representing functions may be adopted in the indexing process. For example, a textual Web page can be indexed both by its textual features and by its link structure features. Thus, the indexing process can also be denoted as:

$$\mathrm{I}_1 : \langle l, A_1 \rangle \to \langle l, (F_1, \mathrm{Op}_1) \rangle$$
$$\mathrm{I}_2 : \langle l, A_2 \rangle \to \langle l, (F_2, \mathrm{Op}_2) \rangle$$
$$\vdots$$
$$\mathrm{I}_N : \langle l, A_N \rangle \to \langle l, (F_N, \mathrm{Op}_N) \rangle$$

I_i, $i = [1, N]$ are representing functions that differ from each other in the choices of attributes A_i, features F_i, and the operations Op_i, $i = [1, N]$.

After indexing, the evaluation of the relationship among the documents is shifted into the evaluation of the relationship among the features. Since the features reside in the space $S = (F, \mathrm{Op})$, where mathematical measures of the relationship (e.g. similarity functions) are defined, the relationship among features are measurable. The indexing process enables us to measure the objects via a set of k features that are extracted and used to represent the objects in the mathematical spaces. In terms of the Web search engines, a reasonable and effective indexing process should make use of the textual content attributes A_t, the structural attributes A_s, and the multimedia content attributes A_m. To put it formally, $\exists\ \mathrm{I}_i \in \mathrm{Id}$, $i \geq 1$, such that, $\mathrm{I}_i : A_i \subset (A_t \cap A_s \cap A_m) \subset A_o \rightarrow (F_d, \mathrm{Op})$.

Definition 7. *Information Need* IN is a collection for all the information obtained directly or indirectly, about a user's requirement for information. IN is a 3-tuple $IN = \langle SR, I, \Psi \rangle$, where

- SR is a user's search request, which can be anything submitted to the search engine by a user;
- I is the information we know about a user in advance. For example, we may know a user's spoken language, fields of interest, preferred Web sites, etc (e.g. by studying cookies);
- Ψ is the deducted information about a user based on both the search request and the information we know about the user, so $\Psi = \Psi(SR, I)$.

A user's information need is more than what is submitted to the search engine in the form of a search request. If we consider a search request as the explicit information we know about a user, then the implicit information about the user's information need contains two parts. The first part is the additional information we know about the user in advance I. The second part is Ψ, which is the information we deduce or infer from both SR and I. In the previous study (Dominich 2001), the deduced information Ψ is considered only deducible from I. We argue that further information can also be deduced from the search request. For example, if a user inputs the search request “java”, we may not know exactly what the user really wants to find out. The user may expect information about the coffee, a programming language, or even a country. However, if we combine this obscure search request with the information we have known about the user, we may understand the search request better. Suppose we know the user is a computer science graduate student. Then there is high possibility that the user is looking for some java programming language relevant information. The above conjecture about a user's search request is an example of the deduced information Ψ.

Definition 8. *Query* Q is a set composed of logical representations for the user information needs IN. A query is produced by applying the representation functions to a user's information need:
$\mathrm{I} : IN \rightarrow Q$; or
$\mathrm{I} : \langle q, A_q \rangle \rightarrow \langle q, (F_q, \mathrm{Op}) \rangle$
Similar to the relationship between objects and documents, a query is the logical/mathematical representation of a user's information need. According to definition 5, a query might only possess a fraction of the attributes of the user's information need, which is represented as F_q. Some query expansion techniques try to overcome the above problem by using several queries simultaneously in order to capture as many features of a user's information need as possible (Lawrence & Giles 2000).

Existing works regarding the goal of Web search consider the problem of evaluating Web pages as finding out the importance of the pages. For example, Arasu et al. (2001) define importance in three ways: pages with textual similarity to the query, popular pages, and pages of high level of URLs. We think the goal of Web search engines is to retrieve those Web pages that are both subjectively relevant to a particular user's (or a set of users') information need, and are objectively important enough for the user (users) to have a look. Important Web pages are not necessarily relevant pages, and relevant pages are not necessarily important pages. Our view of Web page evaluation leads to the following definition of *ranking criteria*:

Definition 9. *Ranking criteria* $C = \{Sub, Obj\}$, where

$Sub = \{s_1, s_2, \cdots, s_U\}, U \geq 1; Obj = \{o_1, o_2, \cdots, o_V\}, V \geq 0$. For example:

$Sub = \{relevance,\ similarity,\ aboutness, \cdots\}$;

$Obj = \{importance,\ authority,\ popularity,\ quality, \cdots\}$.

Ranking criteria is the collection for all the criteria that can be used to evaluate/rank a Web page and it can be divided into two subsets, subjective criteria (Sub) and objective criteria (Obj). In the subjective set, the most common criteria are relevance, similarity, and aboutness. Some researchers think relevance and aboutness are two different terms referring to the same concept (Bruza et al. 2000); other researchers believe that aboutness is distinct from relevance (van Rijsbergen 2005, van Rijsbergen 2004). In this paper, we do not try to differentiate between these two concepts. We just use them as two examples for the subjective ranking criteria. For objective criteria, importance, authority, popularity (Lawrence & Giles 2000), and quality (Cho et al. 2005) are the common criteria being used. All these objective criteria have the meaning of "importance" to some level. Many studies use the same term "importance" but use it in a different conceptual sense. For HITS and PageRank, the importance of a Web page is defined according to the incoming links of the page. In (Beg 2005) the number of times that links having been selected in previous searches is used to define importance. Cho et al. (Cho et al. 2005) defines the quality of a Web page as the conditional probability that an average user will like the page and create a link to the page given that the user discovers the page for the first time.

Definition 10. For $\Theta = 2^C$, the *ranking function* $\mathrm{R}(D, Q, \Theta)$ is a mapping that:

$R : D \times Q \times \Theta \rightarrow RS_Q^\Theta = \left\langle R_{Q,\Theta}^n,\ \leq_{Q,\Theta}^r \right\rangle$,

where RS_Q^Θ is the retrieval set as defined in definition 11. For a given query Q and the ranking criteria Θ, the ranking function produces a subset of documents $R_{Q,\Theta}^n$ from the document corpus D. An order $\leq_{Q,\Theta}^r$ is defined on this subset, so that the resulting RS_Q^Θ is a collection of documents within which an order is defined. The order associated with each document is called the rank value of the document. It is denoted by little case r as $r(D, Q, \Theta)$.

Table 1. Properties Comparison

Property	GF model	UC model	STE model	BYRN model	SE model
fundamental components being defined	X	X	X	X	X
fundamental relationship and	X				
Ranking enabled				X	X
completeness				X	X
media independent		X			X
user-centered		X			X
taking WWW and Web documents characterizations into account					X
multiple ranking criteria enabled					X
capable of handling the dynamics of IR/ Web search engines					

Definition 11. *Retrieval set* RS is a total ordered set of the top n highest ranked documents for a given ranking function and information need: $RS_Q^\Theta = \left\langle R_{Q,\Theta}^n,\ \leq_{Q,\Theta}^r \right\rangle$, where the subscripts Q and Θ indicate

the query and criteria being used. R_Q^n is a subset of the corpus D, and $\leq_Q^r$ is the order/relation defined on the set R_Q^n. The order $\leq_Q^r$ is defined by the ranking function, such that,

- $R_{Q,\Theta}^n \subset D$;
- $\left|R_{Q,\Theta}^n = n\right|$;
- $\forall d_i, d_j \in R_{Q,\Theta}^n$, $i \neq j$, $d_i \leq_{Q,\Theta}^r d_j \Leftrightarrow r\left(d_i, Q, 2^C\right) \leq r\left(d_j, Q, 2^C\right)$. For any two different documents d_i and d_j in the retrieval set, if these two documents satisfy the relation $\leq_{Q,\Theta}^r$ by $d_i \leq_{Q,\Theta}^r d_j$, then the rank value of document d_i is less than or equal to the rank value of document d_j, and vise versa;
- $\forall d_p \in R_{Q,\Theta}^n,\ p = [1, \left|R_{Q,\Theta}^n\right|],\ \forall d_q \in \left(D - R_{Q,\Theta}^n\right),\ q = [1, |D| - n + 1],\ r(d_p, Q, \Theta) \geq r\{d_q, Q, \Theta\}$. All the documents in the retrieval set have rank values that are larger than of those documents in the corpus D that are not being retrieved.

The retrieval set can be expressed in a more expressive way: $RS_Q^\Theta = \{(d_i,\ r(d_i, Q, \Theta))\}$, $i = [1, n]$. It means that the retrieval set is a collection for the pairs of a document and its rank value.

According to the choices of ranking criteria, ranking functions can be further divided into three major types: 1. subjective ranking-criterion based ranking functions, 2. objective ranking-criterion based ranking functions, and 3. the ranking functions that use both subjective and objective ranking-criteria. These three types of ranking functions are expressed as $R_{sub}(\cdot)$, $R_{obj}(\cdot)$, and $R_{so}(\cdot)$ respectively, and are defined as:

- $R_{sub}(Q): D \times Q \times sub \rightarrow RS_Q^{sub}$. The ranking strategy of $R_{sub}(Q)$ function focuses on the content and attributes of individual documents. Search engine users are the final evaluators of this ranking strategy.
- $R_{obj}(\cdot): D \times obj \rightarrow RS^{obj}$. The ranking of documents is independent of a query. In other words, the evaluation of Web documents (objects to be retrieved) is independent of a user's information need. Therefore, search engine users are not the evaluators of the ranking strategy.
- $R_{so}(Q) = g(r(D, Q, sub), r(D, obj)) = g(r_{sub}(Q), r_{obj}(\cdot))$. This ranking function adopts both the objective and the subjective ranking criteria.

Implementing function $g(\cdot)$ with different forms yields different ranking strategies. Three possible forms of function $g(\cdot)$ are:

- $g1(\cdot) = R_{obj}\{R_{sub}(Q)\} = R_{obj}(RS_Q^{sub}) = r_{s,o}$
- $g2(\cdot) = R_{sub}\{R_{obj}(\cdot), Q\} = R_{sub}(RS_Q^{obj}) = r_{o,s}$
- $g3(\cdot) = k_1 \cdot r_{sub}(Q) + k_2 \cdot r_{obj}(\cdot)$, where k_1 and k_2 are two constants.

(The above definition by no means restricts other possible forms for function $g(\cdot)$).

Definition 12. A *Web search engine* is a tuple $\left\langle \mathrm{R}\{[\mathrm{I}(OB, IN)],\ C\},\ RS \right\rangle$ where

- OB is a set for objects;
- IN is the set for user information need;
- I is the representing function;
- R is the ranking function;
- C is the collection of ranking criteria;

- RS is the retrieval set.

Discussion

We compare the performance of the proposed SE model with the other four DO IR models and generalize of the results Table 1. Nine properties are used to evaluate the performance. The first two properties enable a formal model to answer the question "What is IR?" A model is complete if it is ranking-enabled and satisfies the first two properties. The SE model is complete. Properties 5 to 9 concern a model's ability to deal with searching in the Web context. Only the SE model possesses all these five properties. The UC model is media independent and user-centered, but it does not make use of the characteristics of the Web and the Web documents, and does not offer multiple ranking criteria. The last property evaluates a model's ability to formalize the dynamic nature of the Web search engine. A Web search engine is a dynamic system because several interactive processes are involved. Unfortunately, none of the existing formal models is capable of formalizing the dynamics of the Web search engine. Formalizing the dynamics of a Web search engine will be part of future work in this area.

Case Study

We choose PageRank and HITS, two representative Web information retrieval strategies, to test the expressiveness of the new model for Web search engine. PageRank and HITS both are designed especially for Web search instead of some traditional IR problems. To our knowledge, there is still no IR model that can fully explain the retrieval strategies of PageRank and HITS.

HITS

Kleinberg (1999) notices the subjectivity inherent in the notion of relevance that necessitates human evaluation for the quality of a search engine and proposes the HITS method that consistently identifies both "relevant" and "authoritative" Web pages. "Relevance" is achieved by applying traditional text-based searches, and corresponds to the subjective quality of a Web page. In contrast, "authority" depends purely on the number and quality of incoming links, thus corresponds to the objective quality of a retrieved Web page. The partition of the evaluation of Web pages into a subjective part and an objective part reduces the uncertainty inherent in information retrieval problems.

The HITS algorithm can be described by the new model as follows:

1. $\mathrm{I}_{D,t} : OB \to D_t$, $\mathrm{I}_{Q,t} : IN \to Q_t$. Represent objects (Web pages) and user information need as logical/mathematical representations, which are documents and queries respectively. The attributes and features being used are the textual attributes and textual features, which are represented by the subscript t;
2. $R_{rel}(D_t, Q_t, rel) = RS_{Q_t}^{k_1}|_{rel}$. Use traditional text-based searching methods and relevant criterion rel to search the query Q, choose top k_1 documents, and form the root set $RS_{Q_t}^{k_1}|_{rel}$ as the starting point;
3. $RS_{Q_t}^{k_1}|_{rel} \to R_l^{k_2}, |R| > |RS| \Leftrightarrow k_2 > k_1$. Use link attributes and features (subscript l) to obtain the base set R by growing the root set RS to include any page that is pointed to or pointed from a page in RS. Notice that, although RS is a total ordered set, R does not necessarily be an ordered set;
4. $R_{auth}(R_l^{k_2}, auth) = RS_l^{k_3}|_{auth}$, $k_3 \le k_2$. Use the objective authority criterion and the link analysis method to find the most authoritative pages among the base set of relevant documents. The final retrieval set $RS_l^{k_3}|_{auth}$, is a total ordered set. The pages in this set both subjectively satisfy a user's information need, and objectively are the most authoritative one among a great amount of relevant documents found on Web.

PageRank

Brin and Page (1998) view Pagerank as “an objective measure” for Web pages' “subjective idea of importance”. PageRank assumes pages with more important backlinks are more important. Such notion of importance emerges from the topological structure of the Web and is independent of the page content. Different from HITS, PageRank assigns a score of importance to each page at crawl time independent of specific queries. The importance measurement is combined with a traditional information retrieval score at query time. One deficiency of PageRank is “topic-drifting”, which is caused by the ill balance between objective and subjective criteria. PageRank can be formally explained with the new model as:

1. $\mathrm{I}_{D,l} : OB \to D_l$. Represent Web pages as logical/mathematical representations D_l by using the features which can capture the structural characteristic of Web, for example, the link features (subscript l);
2. $\mathrm{R}_{imp} : D_l \times imp \to RS_l^N \mid_{imp}$. Assign a rank to every Web page using *importance*, the objective criterion. N is large enough, standing for the size of the whole Web. The rank for each component of $RS_l^N \mid_{imp}$ is expressed as $r_l^n \mid_{imp}, n = [1, N]$;
3. $I_{D,t} : OB \to D_t$, $I_{Q,t} : IN \to Q_t$. Represent Web pages and a user’s information need as logical/mathematical representation D and Q respectively by using textual features (subscript t).
4. $R_{rel} : D_t \times Q_t \times rel \to RS_{Q_t}^K \mid_{rel}, K << N$. Rank Web pages again by using relevance, the subjective ranking criterion. Normally, according to an individual user's concrete information need, only a sub graph of the whole Web is evaluated at this step. The size of the retrieval set is much smaller than the size of the whole Web. The rank for each component of $RS_{Q_t}^K \mid_{rel}$ is expressed as $r_{Q_t}^k \mid_{rel}$, $k = [1, K]$;
5. $\forall d_i \in (RS_l^N \mid_{imp}) \cap (RS_{Q_t}^K \mid_{rel})$, $R_{s,o}(\cdot) = R_{imp,rel}(\cdot) = g(r_l^n \mid_{imp}^i, r_{Q_t}^k \mid_{rel}^i) = g(r_{imp}^i, r_{rel}^i)$, Combine the objective ranking score with the subjective ranking score to obtain the final ranking score for each object (Web page), and use this final score to return retrieved objects to the user.

Conclusion

This paper begins the process of building a formal model for Web search engines to provide Web search engines with needed formal theory to assist developing future models and techniques. We clarify confusion within the area of IR formal models study by categorizing IR formal models into three major categories. Specific meaning of the term “model” within each category is discussed and generalized. The DO IR models are analyzed to uncover their advantages and inefficiencies for Web IR. In the proposed model, basic components for Web search engines, the relationships that connect basic components and the operations that determine the functionalities of the Web search engines are defined. The new model copes with the specialties of Web, the Web documents, and the Web users. Specifically, the new model incorporates both objective and subjective ranking criteria to handle the abundance problem. The model is also media independent and user-centered. The model proposed has limitations. For example, the interactive nature and the temporal properties of Web search engines have not been covered in the model yet. Improvements to the model are possible that could improve its expressiveness as well as its prediction capabilities. Our plan for future work is to model the temporal nature of Web search engines.

References

1. Nielsen netratings for search engines. (2005). Retrieved November 2005, from http://searchenginewatch.com/reports/article.php.
2. Arasu, A., Cho, J., Garcia-Molina, H., Paepcke, A., Raghavan, S. (2001). Searching the web. *ACM Trans. Internet Techn*, 1 (1), 2-43.
3. Baeza-Yates, R. A., Ribeiro-Neto, B. A. (1999). Modern Information Retrieval. *ACM Press / Addison-Wesley*.

4. Beg, M. M. S. (2005). User feedback based enhancement in web search quality. *Inf. Sci,* 170 (2-4), 153-172.
5. Bianchini, M., Gori, M., Scarselli, F. (2005). Inside pagerank. *ACM Trans. Inter. Tech*, 5 (1), 92-128.
6. Brin, S., Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 30 (1-7), 107-117.
7. Brooks, T. A., 2003. Web search: how the web has changed information retrieval. Inf. Res. 8 (3).
8. Bruza, P., Song, D., Wong, K.-F. (2000). Aboutness from a commonsense perspective. *JASIS,* 51 (12), 1090-1105.
9. Cho, J., Roy, S., Adams, R. E. (2005). Page quality: in search of an unbiased web ranking. *SIGMOD '05*, 551-562.
10. Crestani, F., Lalmas, M. (2001). Logic and uncertainty in information retrieval, *Lectures on information retrieval*, 179-206.
11. Cutler, M., Shih, Y., Meng, W. (1997). Using the structure of html documents to improve retrieval. *USENIX Symposium on Internet Technologies and Systems*.
12. Dominich, S. (2001). On applying formal grammar and languages, and deduction to information. *Proceedings of the ACM SIGIR MF/IR*, 37-41.
13. Egghe, L., Rousseau, R. (1998). A theoretical study of recall and precision using a topological approach to information retrieval. *Inf. Process. Manage.* 34 (2-3), 191-218.
14. Everett, D. M., Cater, S. C. (1992). Topology of document retrieval systems. *JASIS* 43 (10), 658-673.
15. Faloutsos, C., Barber, R., Flickner, M., Hafner, J., Niblack, W., Petkovic, D., Equitz, W. (1994). Efficient and effective querying by image content. *J. Intell. Inf. Syst.* 3 (3-4), 231-262.
16. Fuhr, N. (1992). Probabilistic models in information retrieval. *Comput. J.* 35 (3), 243-255.
17. Goncalves, M. A., Fox, E. A., Watson, L. T., Kipp, N. A. (2004). Streams, structures, spaces, scenarios, societies (5s): A formal model for digital libraries. *ACM Trans. Inf. Syst*. 22 (2), 270-312.
18. Grossman, D. A., Frieder, O. (2004). *Information Retrieval: Algorithms and Heuristics*. Springer.
19. Jain, R. (1995). World-wide maze. *IEEE MultiMedia* 2 (2), 3.
20. Kherfi, M. L., Ziou, D., Bernardi, A. (2004). Image retrieval from the world wide web: Issues, techniques, and systems. *ACM Comput. Surv.* 36 (1), 35-67.
21. Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *J. ACM* 46 (5), 604-632.
22. Lawrence, S., Giles, C. L. (2000). Accessibility of information on the web. *Intelligence*, 11 (1), 32-39.
23. Meng, W., Yu, C., Liu, K.-L. (2002). Building efficient and effective metasearch engines. *ACM Comput. Surv.* 34 (1), 48-89.
24. Rolleke, T., Tsikrika, T., Kazai, G. (2003). A general matrix framework for modeling information retrieval. *Proceedings of the ACM SIGIR MF/IR*, 1-11.
25. Sheridan, P., Braschler, M., Schauble, P. (1997). Cross-language information retrieval in a multilingual legal domain. *ECDL '97*, 253-268.
26. Silverstein, C., Marais, H., Henzinger, M., Moricz, M. (1999). Analysis of a very large web search engine query log. *SIGIR Forum,* 33 (1), 6-12.
27. Tague, J., Salminen, A., McClellan, C. (1991). Complete formal model for information retrieval systems. *SIGIR '91*, 14-20.
28. Tsikrika, T., Lalmas, M. (2002). Combining web document representations in a bayesian inference network model using link and content-based evidence. *Proceedings of the 24th BCS-IRSG European Colloquium on IR Research,* 53-72.
29. van Rijsbergen, C. J. (2004). *The Geometry of Information Retrieval*. Cambridge University Press.
30. van Rijsbergen, C. J. (2005). A probabilistic logic for information retrieval. *ECIR*. pp. 1-6.