Rational Research Model for Ranking Semantic Entities

Wang Wei, Andrzej Bargiela, *Member, IEEE,* and Payam Barnaghi, *Member, IEEE*

Abstract—Ranking plays important roles in contemporary information search and retrieval systems. Among existing ranking algorithms, link analysis based algorithms have been proved as effective means for ranking documents retrieved from large-scale text repositories, such as the current Web. Recent development in semantic Web raises considerable interests in designing new ranking paradigms for various semantic search applications. While ranking methods in this context exist, they have not gained much popularity. In this article we introduce the idea of the "Rational Research" model which reflects search behaviour of a "rational" researcher in a scientific research environment, and propose the RareRank algorithm for ranking entities in semantic search systems, in particular, we focus on elaborating the rationale and implementation of the algorithm. Experiments are performed using the RareRank algorithm and the results are evaluated by domain experts using popular ranking performance measures. A comparison study with existing link-based ranking algorithms reveals the preponderance of the proposed method.

Index Terms—Ranking, Ontology, Semantic Search, Rational Research, RareRank Algorithm.

1 Introduction

Most of the contemporary information search and retrieval systems present results in a ranked list to users by employing certain ranking algorithms or functions. Among various ranking algorithms, link analysis based ones have been proved as effective methods to rank information retrieved from large scale document repositories, such as the Web. Some of the link analysis based algorithms attempt to simulate human search behaviour, for example, the PageRank algorithm [1], [2] assumes a "random surfer model" which imitates human search behaviour in a hyper-linked environment; the Hypertext Induced Topic Selection (HITS) algorithm [3] calculates "authority" and "hub" values that reflect users' intuition; in scientific research, a citation link established between two documents is regarded as incorporating human cognitive judgement of quality: ranking method such as "Autonomous Citation Indexing" [4] ranks publications based on number of citations that have been made to them. These algorithms have been implemented in popular search engines that are being used by millions of users. To balance the search quality and relevance, real retrieval systems usually resort to sophisticated parameter tuning techniques to integrate the link based ranking and relevance judgement (content analysis) scores to provide final ranking results.

We presents the "Rational Research" model, a link analysis based ranking algorithm, in the context of scientific research. The idea behind the model is that entities in knowledge base such as documents, authors, journals and conferences, together with topics in terminological ontology, can reasonably simulate an environment in which researchers explore scientific publications related to their research interests. The produced ranking naturally combines the link information (e.g., a citation between two publications), and the content information (e.g., provided by the links between document-topic and topic-topic). The model provides an appropriate basis for ranking various types of entities and clearly can be generalised into other domains.

The rest of the paper is organised as follows. Section 2 reviews some of the representative link analysis based ranking algorithms in the information retrieval community, and entities ranking methods developed in the semantic Web community. In Section 3 we elaborate the "Rational Research" model, in particular, its justification, principle, and implementation details (The algorithm is referred to as "RareRank"). Section 4 explains the experiments conducted using the proposed RareRank algorithm. In Section 5 we define the evaluation measures, and present the evaluation results. Moreover, the results are compared with those generated using two representative algorithms, i.e., the original PageRank [1] and ObjectRank [5], [6] under same experimental settings. Section 6 concludes the paper and discusses the future work.

2 Related Work on Ranking

Ranking has become indispensable in modern information retrieval systems, which strive to find quality and relevant results from extraordinary huge document repositories (e.g., the Web). Classic information retrieval (IR) models such as the Vector Space model and the Probabilistic model are effective for finding

- *Wang Wei and Andrzej Bargiela are with the School of Computer Science, University of Nottingham (Malaysia Campus). Payam Barnaghi is with the Centre for Communication Systems Research (CCSR) at the University of Surrey.*
  E-mail: wang.wei@nottingham.edu.my, abb@cs.nott.ac.uk, p.barnaghi@surrey.ac.uk

relevant information, and they also provide means to compute content-based ranking for the results. However, the content-based ranking paradigm has only achieved modest performance. The BM25 weighting scheme [7] offers means for parameter tuning to achieve reasonable ranking results and has been utilised by many IR research groups, however, the tuning process is tedious and the derived parameter setting is not likely to work for all document corpora. During the past decades, Latent semantic models such as Latent Semantic Analysis (LSA) [8], and its probabilistic variants probabilistic Latent Semantic Analysis (pLSA) [9] and Latent Dirichlet Allocation (LDA) [10], [11] have been formulated as advanced content analysis techniques. They have been proved as being effective for document modelling and dimensionality reduction by large number of published works [8], [9], [10], however, their computation complexity and inscalability have restrained them from being used for the purpose of ranking in real retrieval systems. As the number of documents becomes unprecedentedly large, it is extremely difficult for search engines to choose just tens of quality documents out of millions with content-based ranking paradigm only.

In this section we first provide a review on link analysis based techniques, in particular the original PageRank which has inspired emergence of many variants. We then discuss some of the representative variants due to their close relatedness to our work and point out their limitations. To make our discussion more complete, we also present some ranking methods used in recently developed semantic search systems and retrieval systems for scientific publications.

2.1 Link Analysis

Link analysis refers to a broad range of techniques for solving specific ranking problems by exploiting link structures, such as links between actors in social networks [12], citation links in scholarly publications [13], [14], [4], and hyperlinks among Web pages [1], [3]. Due to their intuitive and reasonable assumptions, and superior practical performance, these techniques have become dominant ranking schemes deployed in many contemporary Web search engines and various vertical search engines for locating authority and quality documents.

In Social Network Analysis, centrality (i.e., degree centrality, betweenness centrality, and closeness centrality) and prestige (degree prestige, proximity prestige, and rank prestige) analysis are two important means to identify important or prominent actors in a social network [12]. In fact they have a similar idea as the later developed link analysis based methods for ranking documents. In the domain of scientific research, the co-citation analysis [13] and bibliographic coupling [14] are two popular approaches to calculate similarities between documents. The autonomous citation indexing [4] enables automatic algorithmic extraction and grouping of citations from publically available scientific publications, and facilitates browsing and retrieval. In 1998, emergence of two ranking techniques, PageRank [1] and HITS [3] immediately attracted the IR community's attention. The PageRank algorithm perhaps is the most important underpinning technology for Google, the dominant search engine giant. The HITS algorithm [3] has been used to identify authority Web pages on the Web, and hub documents in publications [15].

2.2 PageRank

The original PageRank algorithm is described in [1], and rational of the algorithm can be explained using the random walk and theory of Markov Chain [2], [16]. A random surfer visits Web pages by following the (hyper)links among them, and the process can be modelled as a Markov Chain with one state for each Web page. A Markov chain is characterised by a stochastic matrix A which has the property as shown in Equation (1).

$$\forall i, \sum_{j=1}^{N} P_{ij} = 1 \tag{1}$$

where $\forall i, j, P_{i,j} \in [0, 1]$. A key property of a stochastic matrix is that it has a principal left eigenvector corresponding to its largest eigenvalue 1 [17], [16]. An important property of a Markov Chain is that for any starting point, the chain will converge to the stationary distribution as long as the transition probability matrix A obeys two properties, Irreducibility and Aperiodicity [17]. The invariant probability distribution and transition matrix thereby satisfy Equation (2).

$$\vec{\pi} P = \vec{\pi} \tag{2}$$

where $\vec{\pi}$ is the stationary probability distribution of a Markov Chain. The PageRank values of all Web pages are essentially the invariant probability distribution of Markov Chain characterised by the transition probability matrix constructed from the Web graph with some rules. To ensure the probability transition matrix of the Web page graph satisfy the irreducibility and aperiodicity properties, the dampening factor d is added into the rank propagation process and the idea of teleport operation [16] is introduced. The resulting transition probability matrix is guaranteed to satisfy the properties of irreducibility and aperiodicity. PageRank value of a Web Page is represented in Equation (3).

$$P_r(i) = (1 - d) + d \sum_{j=1}^{N} A_{ji} P_r(j) \tag{3}$$

where $P_r(i)$ is the PageRank of a Web page i, A_{ji} is the transition probability from node j to i. Computation of the PageRank can be done using the power iteration method [16] which terminates when the PageRank vector converges.

2.3 Variants of PageRank

The teleport operation in the original PageRank is uniform, i.e., from a Web page, the probabilities of transition to each of the other pages are same. This setting has been criticised since the uniform treatment of transition is often unrealistic. There has been research towards the topic sensitive PageRank [18], in which a number of scores of prominence of a page with respect to various topics are computed. At the query time, these scores are combined based on the query to form a composite PageRank score. However, the method involves large amount of pre-processing with respect to a number of flat topics, and the relationships between topics are not discussed.

Richardson and Domingos proposed an idea of "intelligent surfer", who probabilistically hops from page to page, depending on the content of the pages and the query terms the surfer is looking for [19]. Based on the original PageRank, the method performs a word-matching between query terms and the linked documents (although the authors claim that more sophisticated content analysis can be applied), and set the transition probability between documents proportional to its relevance scores. However, it is likely to ignore those documents which are highly relevant to the query while not linked to the current document.

ObjectRank [5], [6] is another variant of PageRank developed for searching and ranking entities in databases of bibliographic information. ObjectRank introduces and distinguishes between the concepts "authority transfer schema graph" and "authority transfer data graph". Such modelling is similar to ours in which "authority transfer schema graph" corresponds to schema ontology and "authority transfer data graph" corresponds to knowledge base in our work. The idea of ObjectRank also has close connection to ranking entities in semantic search systems since it models different types of nodes in the authority transfer graphs. One of the major limitations is that is does not incorporate the topic entity, which is important in the domain of scientific publication. Except keyword matching, the algorithm does not integrate content analysis for documents. Moreover, calculation of the keyword-specific and global ObjectRank scores uses same information repeatedly and might generate results far from optimal.

2.4 Ranking in Semantic Search Systems

Different from traditional text-based information retrieval systems which exclusively retrieve and rank documents, semantic search systems need to retrieve and rank entities of various types. Usually semantics of links among entities are defined in schema ontologies (e.g., through the domain and range constructs in RDF/S or OWL languages[1]). Ranking algorithms are required to take into account the distinction between semantic links and hyperlinks.

The Swoogle[2] semantic search engine [20] focuses on retrieving and ranking ontologies on the Web using the OntoRank algorithm [21]. OntoRank differentiates four types of semantic links (i.e., imports, uses-term, extends, and asserts) among SWDs, and ranking score of a SWD is computed using a PageRank-like algorithm in which weights of the different semantic links are reflected. ReConRank [22] is the underlying ranking algorithm for SWSE[3], a semantic search engine for searching and retrieving entities and simple knowledge [23]. The algorithm can be seen as performing a three-step computation for entity prioritisation: in the first step, data crawled from the Web is transformed to directed labelled graph, and ResourceRank algorithm is applied to compute ranking scores for resources. The second step extracts context graphs using the provenance of the data to compute ContextRank scores. In the final step, a resource-context graph is derived based on some predefined rules, and the two algorithms are integrated to produce ReConRank scores, which reflects importances of a resource itself as well as its context.

2.5 Ranking in Scientific Research

Our investigation on ranking in the domain of scientific research reveals that there are two dominant approaches: content-based and citation-based ranking. Some of the large online digital libraries employ content-based ranking strategies, such as the IEEE *Xplore*[4] and ACM[5] digital libraries. Another popular search engine for retrieving publication is Google Scholar[6] whose ranking strategy is based on "weighing the full text of each article, the author, the publication in which the article appears, and how often the paper has been cited in other scholarly literature". Although the details of the ranking algorithm in Google Scholar is unknown, it can be perceived that it is a hybrid approach that combines both content and citation-based features. Scirus[7] is another search engine in the domain of scientific research which employs a similar ranking strategy. In the $CiteSeer^X$ search engine, the results are primarily ranked based on the number of citations. A recent work uses variable-strength conditional preferences [24] with a Description Logic knowledge base to ranking objects. The approach allows for formulating complex user queries with rich semantics. However, formulation of these complex queries is not trivial, and in a sense it does not provide a real ranking scheme because the ranking is only based on metadata, in particular, conditional preferences satisfiability, in responding to queries.

[1] RDF/S and OWL are two languages proposed by the World Wide Web Consortium (W3C) to develop ontologies on the semantic Web (see http://www.w3.org/standards/semanticweb/).

[2] http://swoogle.umbc.edu/

[3] http://swse.deri.org/

[4] http://ieeexplore.ieee.org/

[5] http://portal.acm.org/

[6] http://scholar.google.com/

[7] http://www.scirus.com/

3 Rational Research Model

An ideal ranking function would be the one that defines a natural and optimal combination of relevance and quality scores. The classic IR models rank documents exclusively based on content (relevance), while the link analysis based methods emphasise link structures (quality). In fact many of the retrieval methods derive ranking scores using combination of both relevance and quality through sophisticated parameter tuning or learning process.

We proposed a model called "Rational Research" (the corresponding algorithm is referred to as "RareRank") which simulates the process that researchers search and explore scientific literature. The basic idea behind the model is summarised as follows. First a knowledge base in a research domain (consisting of instances such as publication, author, and journal or conference) is represented as a directed and labelled graph. Then a domain topic ontology is plugged into the graph. Weights of the links between topics in the topic ontology and documents in the knowledge base are established according to their similarity values (The weights are calculated using the Latent Dirichlet Allocation (LDA) model [10], [11] and we reuse the program developed in [25]). The entire graph (labelled, directed, and weighted) can be used to simulate an environment in which a researcher explores and searches for publications. Computation of the ranking scores is based on the principle of convergence of a Markov Chain, and the transition probability matrix is constructed based on two sets of transition rules (see Section 3.2 and 3.3 for details). The derived ranking score naturally integrates both relevance (e.g., using the domain topic ontology) and quality (e.g., using the citation links). The model still needs parameter tuning, however, the tuning procedure is intuitive and simple, and only involves setting weights of links in the schema ontology (the size of the schema is very small compared to knowledge base, see Section 4.2).

3.1 Model Justification

When people browse and search for publications, they do not always follow explicit links such as citation or "published-in" (a paper is published in a journal). In many situations, they follow some invisible or indirect links between documents with similar or closely related topics relevant to their research interests. Such links are formed based on human cognitive processing of information and the research environment, however, they are not modelled in many previous link analysis based methods. In Bender *et al*'s work of exploiting social relations for result ranking [26], similarity between tags is computed directly using the Dice coefficient. In contrast, in "Rational Research", links between documents are modelled indirectly by terminological ontologies. Such modelling has some advantages: in our ranking method, similarity value between documents need not to be explicitly calculated; using topics to model associations among documents is generally superior to using words based on classic IR methods [9], [10], [11]; more importantly, we could navigate from one document to others indirectly using the established links between topics and documents. This naturally simulates an important way of searching publications in research environment where topics (or subjects) play fundamental roles and searchers normally have certain level of understanding about the research domain.

3.1.1 Berrypicking

The procedure of searching for scientific information is also described by the "Berrypicking" model [27], which summarises typical search behavior of researchers. The model assumes that a user has several different search strategies such as "footnote chasing", "citation searching", "journal run", "area scanning", "subject searches", and "author searching" (details can be found in [27]). The "Rational Research" model in fact accommodates all of the above searching strategies which are modelled by the different relationships between classes in the schema ontology and between entities in the knowledge base. Mapping between the search strategies in "Berrypicking" and the modelling relationships (see Figure 1) in "Rational Research" is listed as follows.

- "footnote chasing" ↦ "cited"
- "citation searching" ↦ *not modelled*
- "journal run" ↦ "publish" and "publishedIn"
- "area scanning" and "subject searches" ↦ "hasTopic", "isTopicOf", "narrower", "broader", and "related"
- "author searching" ↦ "write" and "isWrittenBy"

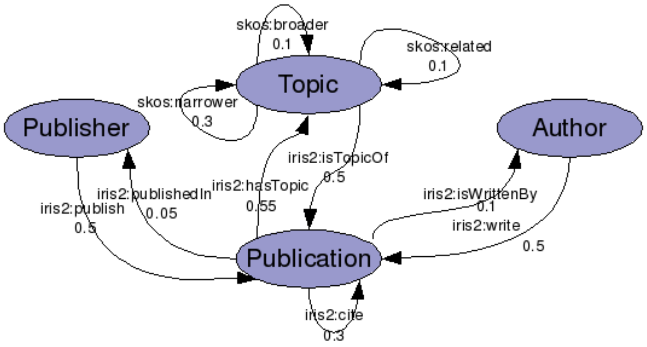


Fig. 1. Relationships and transition probabilities defined in the ontology schema graph

The reason that we choose not to consider the "citation searching" is because that the impact of citing papers is not as significant as the cited ones. However, it can be easily incorporated into our model.

3.1.2 Simulation of Research Environment

Beside document and topic entities, there are other entities of various kinds in the research domain that are of interest to searchers, such as authors, publishers (e.g., conference and journal). These entities facilitate people to search and find published literature in the real world, for example, by issuing queries to a search engine or browsing categories in digital libraries, following citation links, browsing conference proceedings and journal issues, and searching for publications in authors' publication list. Intuitively, existence of the entities and the semantics of the relationships among them simulate a typical scientific research environment.

3.1.3 Reflection of Searcher Behaviour

The proposed model also intuitively reflects human's cognitive processing of information and represents a justifiable means for modelling researchers' searching activities and behaviour.

An exemplar scenario helps to depict our perception: a researcher uses a search engine (the IEEE *Xplore*, ACM, or Google Scholar) to find documents related to his research interests. If he has a clear topic in his mind (normally it is the case), the documents listed on the top or first few pages will have higher probability of being downloaded and read. If he thinks the topic is too general, he might figure out more specific or narrower topics, and reissue queries to retrieve another set of documents. Suppose after he reads the abstracts of some documents of interest; he observes that the authors use some techniques from other research fields (related topics) with which he is not familiar. He probably will search for documents with those related topics. If he is familiar with the topic of interest, he is likely to follow the links between publications and journals, or authors, and start to browse the conference proceedings, different issues in a journal, or publication list of the authors. However, to our knowledge, this kind of behaviour is not modeled in the PageRank, HITS, and other ranking methods.

With the addition of the terminological ontology into the knowledge base, relationships among entities (conference, journal, topics, authors, etc) are enriched. More importantly, the different types of links (either explicitly or implicitly) incorporate semantics of human behaviour. For the reasons, we name the proposed model as "Rational Research" model. It assumes that a researcher will make rational choice as opposed to "random walk" in the original PageRank model [1].

3.1.4 Authority Flow

The proposed method can also be explained in terms of "authority" flow. In a pure citation graph as in the original PageRank or HITS, "authority" only flows through the single type of links (i.e., citations), while in our method, "authority" flows through richer types of relationships, which enables ranking values of various entities to reinforce each other. For example, a document entity with high ranking values would contribute more "authority" to its surrounding entities (e.g.,its authors), consequently, their rankings would be promoted. Authoritativeness of a document is not only dependent on how many citations have been made to it, but also how prominent its authors are, and how related it is to the topics to queries. Similarly, ranking of a journal is promoted if it has published many high quality documents. Compared to the ranking algorithms implemented in most of the search engines, the proposed model combines the relevancy and quality of documents in a more natural way.

Another advantage is that it is able to promote the presence of newly written while highly relevant documents with regard to queries. This partially solves the problem of pure citation analysis based ranking as pointed out in [4]. Besides ranking of documents, rankings of researchers (authors), conferences and journals could also be provided.

3.2 Transition Probability in Ontology

PageRank values are in fact invariant probability distribution of an irreducible and aperiodic Markov Chain which is defined by a stochastic matrix [2], [16] constructed from the Web graph. To make sure the chain is both irreducible and aperiodic, a complete set of outgoing links from each Web page to all others is added. In other words, from each node, there is a probability, called the teleport, to reach all the other nodes in the Web page graph.

Computation of RareRank is based on the same principle, however, the major difference between RareRank and PageRank is the definition of the transition matrix. In the RareRank, there are two transition graphs: the ontology schema graph and the knowledge base graph. The schema graph designates the relations between ontological classes and their transition weights. The knowledge base graph consists of instances (or entities) and their relationships instantiated from the schema ontology. Weight of a relation from an instance i_a to another instance i_b is determined by the weight of the relation between the classes of i_a and i_b defined in the schema graph, how many instances of the same type as i_b that i_a links to, as well as strength of the association between instances.

Before we discuss the transition probability issues in the ontology schema and knowledge base graphs, we give definitions of four terms related to the teleport operation which underpins the RareRank model.

Definition 1: **Full Teleport Probability**: is the probability to initiate a teleport operation when a class has no outgoing links in the ontology schema, or an instance in the knowledge base has no outgoing links.
It is denoted by p_f^t and has value of 1. Note that the probability of teleport from one instance is $1/N$, where N is the total number of entities.

Definition 2: **Base Teleport Probability**: is the probability to initiate a teleport operation when a class has outgoing links in the ontology schema (then an instance of the class in the knowledge base possibly has outgoing links).

It is denoted by p_b^t and is set to $1 - d$, where d is the dampening factor.

Definition 3: **Schema Imbalance Teleport Probability**: is the probability to initiate a teleport operation when sum of weights of the outgoing links of a class is less than 1.
It is denoted by p_{si}^t. In this case, the value of the difference between 1 and total weights of the outgoing links of a class will be transferred for teleporting.

Definition 4: **Link Zero-instantiation Teleport Probability**: is the probability to initiate a teleport operation when a predicate is defined in schema, but not instantiated in the knowledge base.
It is denoted by p_{zi}^t. If a predicate of a class is not instantiated in the knowledge base, the weight of the predicate is transferred for teleporting.

Therefore, if a class in the schema has outgoing links, then the probability of teleport operation is $p^t = p_b^t + p_{si}^t + p_{zi}^t$.

3.2.1 Transition Probability in Schema Graph

The schema of the IRIS2 publication ontology [28] is translated into a directed and weighted graph. The direction between two nodes in the schema graph is defined as from the domain to the range of a relation in the schema ontology and the weights of the graph are configurable parameters. The notations used in the schema graph are shown below.

- O - publication schema ontology graph;
- N_C - number of classes defined in O;
- C - set of classes defined in O, $C = \{c_i | c_i \in C, 0 < i \leq N_C\}$;
- N_P - number of predicates (relations) defined in O;
- P - set of predicates defined in O, $P = \{p_j | p_j \in P, 0 < j \leq N_P\}$;
- p_j - the jth predicates;
- $w_{p_{j,c_i}}$ - weight of a predicate p_j whose domain is the class c_i, $w_{p_{j,c_i}} \in [0, 1]$;
- $|OL_{c_i}|$ number of outgoing links (predicates) from class c_i.

Figure 1 shows the schema graph and one of the typical settings of predicate weights defined in our experiment. Weights of the predicates in the schema graph are manually set (the number of weights need to be set is small, in our case is only 10. See Section 4.2 for more discussion on the initial setting of predicate weights) prior to the computation of transition probability matrix for the graph transformed from the knowledge base. The weights reflect the semantics of the domain [5] and user's preferences, for example, when a user is visiting an Publication node, he has 0.1 probability to traverse to the Author node, 0.05 to the Publisher, 0.55 to the Topic, and 0.3 to another Publication node (through "cite" relation).

In Figure 1, "skos" is the prefix for the namespace of SKOS ontology, and "iris2" is the namespace prefix of the IRIS schema ontology. There are 4 classes and 10 predicates in the graph:

- Topic - iris2:Topic;
- Author -iris2:CSResearcher;
- Publisher - intersection of iris2:Conference and iris2:Journal;
- Publication - iris2:Publication, which is the super class of iris2:InProceedings, iris2:Article, etc.

3.2.2 Transition Rules in Schema Graph

We define a number of rules for transition probability between classes in O.

Definition 5 (Probability Transition Rules in Schema):
Let the dampening factor be a constant d.

- **Rule 1**: If a class does not have any outgoing links (predicates), then the teleport operation is initiated with probability of 1;
- **Rule 2**: If the sum of transition probabilities from one class c to all other classes is greater than 0[8], $\sum_j^{|OL_{c_i}|} w_{p_{j,c_i}} = 1$, then the teleport is initiated with probability of $1-d$, i.e., the Base Teleport Probability p_b^t;
- **Rule 3**: If the sum of transition probabilities from one class to all other classes is less than one, $\sum_j^{|OL_{c_i}|} w_{p_{j,c_i}} \in (0, 1)$, then the teleport probability is increased by value of $d(1-\sum_j^{|OL_{c_i}|} w_{p_{j,c_i}})$, i.e., the Schema Imbalance Teleport Probability p_{si}^t;

In RareRank, a typical value of d is set as 0.95, as opposed to the value of 0.85 in the original PageRank [1], [2] (ObjectRank [5] also uses 0.85). The reason we adopt a smaller probability for the teleport operation is that there is less "randomness" in the research domain compared to general Web search. The number $1 - d$ is to ensure the instance graph (translated from the Knowledge Base) is fully connected and does not get trapped in cycles. The Rule 1 and 2 are straightforward as they are similar to those defined in PageRank [16]. The Rule 3 designates that if the sum of probabilities from one class in the schema is less than one, than the difference between the sum and 1 will be transferred to teleport operation. For example, when a user is at an Author node, he has 0.5 probability to traverse to the author's publication nodes; the rest of 0.5 is not specified and thus will be transferred for teleporting. In some situations, the relations defined between two classes in the schema might not be instantiated in the knowledge base. As described in the Section 3.3, the rules defined for ontology schema have to be used in combination with those defined for knowledge base in order to construct a transition probability matrix that is both irreducible and aperiodic.

3.3 Transition Probability in Knowledge Base

The knowledge base consists of all the instances I of the classes C and predicates P defined in O. The transition probability matrix is computed based on this graph

[8] If the sum is greater than 1, normalisation is needed to ensure the sum equals to 1.

conforming to the rules defined for O. The notations are listed below.

- K - knowledge base graph;
- I - all instances defined in K whose types are classes C in O;
- $i(c)$ - an instance of the class c;
- IP - all predicate instances instantiated in K;
- N - number of instances in K;

3.3.1 Transition Rules in Knowledge Base

We define rules for transition probability between instances of classes C in the graph of K as follows.

Definition 6 (Probability Transition Rules in KB): Let the knowledge base graph K conform to the ontology schema graph O,

- **Rule 4**: If an instance does not have any outgoing links to any other instances in the K, then the teleport operation of the instance is initiated with probability of $1/N$ to any other instances.
- **Rule 5**: If an instance has one or more outgoing links, then the Base Teleport Probability for the instance is set to $p_b^t = (1-d)/N$.
- **Rule 6**: If the sum of transition probabilities from a class c_i to all other classes is less than one, $\sum_j^{|OL_{c_i}|} w_{p_{j,c_i}} \in (0,1)$, then the teleport from an instance of c_i is increased by probability of $d(1 - \sum_j^{|OL_{c_i}|} w_{p_{j,c_i}})/N = p_{si}^t/N$.
- **Rule 7**: If an instance of a class c_i does not instantiate one or more predicates defined in the ontology schema, the teleport from the instance is increased by probability of $d\sum_{p_j \notin IP}^{|OL_{c_i}|} w_{p_{j,c_i}}/N$.
- **Rule 8**: If a predicate p_{j,c_i} is present in K, then count the number of occurrence of p_{j,c_i}, the transition probability is defined as $dw_{p_{j,c_i}}/|p_{j,c_i}|$.

In Rule 8, $|p_{j,c_i}|$ is the number of times that the predicate p_{j,c_i} is instantiated in K.

3.3.2 Transition Probability Computation

Following the rules 1 to 8 defined above, value of one cell in the transition probability matrix, i.e., the transition probability from instance i to instance j, can be calculated using Equation (4).

$$A_{ij} = \begin{cases} 1/N & \text{no out-links} \\ A_{ij}(1) + A_{ij}(2) + A_{ij}(3) + A_{ij}(4) & \text{else} \end{cases} \tag{4}$$

where

$$A_{ij}(1) = (1-d)/N \tag{5}$$

$$A_{ij}(2) = d(1 - \sum_{k,p_{k,c_i} \in P}^{|OL_{c_i}|} w_{p_{k,c_i}})/N \tag{6}$$

$$A_{ij}(3) = d \sum_{k,p_{k,c_i} \in P, p_{k,c_i} \notin IP}^{|OL_{c_i}|} w_{p_{k,c_i}}/N \tag{7}$$

$$A_{ij}(4) = dw_{p_{j,c_i}}/|p_{j,c_i}| \tag{8}$$

In Equation (4), if the instance i has no outgoing links (also referred to as a "rank sink" in PageRank [1]), then the computation is straightforward. If i has one or more outgoing links, the computation involves four terms. The first three terms in fact compute the total probabilities for teleport operation, $p^t = p_b^t + p_{si}^t + p_{zi}^t$. The computation iterates over all instances in the K. The last term is the probability of jumping from i to j, assuming the jumping is uniform. It can be modified to accommodate the non-uniform case by adding a normalised weight for the instances of the predicate. For example, if a publication node links to few topic nodes, we can add weights for each of the links, which are similarity values between the publication and topics. Then the term jumping probability, p^j can be written using Equation (9).

$$p^j = d \cdot w_{p_{j,c_i}} \cdot \frac{sim(i,j)}{\sum_k sim(i,k)} \tag{9}$$

where the function $sim(i,j)$ is the similarity value (e.g., Cosine similarity) between instances i and j. The terms $sim(i,j)/\sum_k sim(i,k)$ is a normalised weight for the predicate.

Following the above discussion, computation of the transition probability matrix A can be decomposed to computation of two matrices: the teleport matrix A_t and jumping matrix A_j as shown in Equation (10).

$$A = A_t + A_j \tag{10}$$

A_t is a $N \times N$ matrix in which each row has the same value. In the jumping matrix, if an instance has links to another instance, then the cell is set to $dw_{p_{j,c_i}}/|p_{j,c_i}|$, otherwise 0. With the rules, the sum of each row in A is guaranteed to be 1, and A is both irreducible and aperiodic. Therefore, the probability vector containing ranking values for entities in the knowledge base is guaranteed to converge to its invariant probability distribution.

3.3.3 Algorithm for Transition Matrix

We have developed an algorithm for computing each row of the probability transition matrix. Assuming that there are N entities in the knowledge base K to be ranked, running the algorithm N times generates the probability transition matrix. Each entity in K is assigned a unique ID, and the algorithm takes it as a parameter and returns a transition probability row.

The Algorithm starts with constructing the ontology schema and knowledge base graphs O and K. The dampening factor d and all $w_{p_{j,c_i}}$ values can be customised according to user′ preferences. Lines 3 to 7 compute the full teleport probability p_f^t since the instance has no outgoing links. Lines 9 to 12 compute the schema imbalance teleport if sum of the outgoing predicates weights is less than one for the class of an

Algorithm 1 Computing Row of Transition Probability Matrix

```
Require: $O$, $K$, $C$, $P$, $d$, all $w_{p_{j,c_i}}$.
Ensure: Probability transition matrix row $M[i]$ using "Rational Research" model.
 1: $p^t = 1 - d$; $M[i] = 0.0$;
 2: get instance $i$'s class $c_i$ from $O$, and retrieve all outgoing predicates $p_{j,c_i}$ and their weights $w_{p_{j,c_i}}$, and save them into a weight vector $V_{sp}$;
 3: if $V_{sp}$ is empty then
 4:     for $j = 0; j < N; j++$ do
 5:         $M[i][j] = 1/N$;
 6:     end for
 7:     return $M[i]$;
 8: else
 9:     if $\sum_j V_{sp}[j] < 1$ then
10:         $p^t_{si} = d(1 - \sum_j w_{p_{j,c_i}})$;
11:         update teleport probability, $p^t = p^t + p^t_{si}$;
12:     end if
13:     for each predicate $p_{j,c_i}$ in $V_{sp}$ do
14:         count the number of times $|p_{j,c_i}|$ the predicate is instantiated in $K$;
15:         if $|p_{j,c_i}| = 0$ then
16:             $p^t_{zi} = dw_{p_{j,c_i}}$;
17:             update teleport probability, $p^t = p^t + p^t_{zi}$;
18:         else
19:             $w^{'}_{p_{j,c_i}} = w_{p_{j,c_i}}/|p_{j,c_i}|$;
20:             save $w^{'}_{p_{j,c_i}}$ into a hashtable $V_{ip}$;
21:         end if
22:     end for
23:     for each instance $j, j \neq i$ in $K$ do
24:         for all predicate type between $i$ and $j$ do
25:             lookup jumping probability $w^{'}_{p_{j,c_i}}$ from $V_{ip}$;
26:             $M[i][j] = M[i][j] + w^{'}_{p_{j,c_i}}$;
27:         end for
28:         $M[i][j] = M[i][j] + p^t/N$;
29:     end for
30:     return $M[i]$;
31: end if
```

instance. The total probability of the teleport operation p^t is incremented accordingly. Lines 13 to 22 compute the link zero-instantiation teleport probability p^t_{zi}, and at the same time, calculate the jumping probability for each type of predicate links which are then saved into a hashtable. The jumping probabilities for an instance to all the other linked instances with the same type (i.e., they are instances of the same class) are uniform. As stated earlier, this can be easily extended to a biased distribution using Equation (9). Lines 23 to 30 calculate the transition probability row $M[i]$ values for the instance i. Each element in the row represents the transition probability from i to all other instances j in the knowledge base. Its value is the sum of the teleport probability for the row and jumping probability.

Time complexity of the RareRank is similar to the original PageRank. In the original PageRank, computation of the transition probability matrix scans all the nodes in the graph and computes the transition probability of one node to all the others. Here we do not try to optimise the computation and simply assume the running time is $O(n^2)$. In RareRank, the computation of the transition probability matrix scans all the nodes according to the rules twice: in the first round the algorithm updates the teleport value, and in the second round, RareRank updates the transition probability of each node to all the others. Therefore, the time complexity of RareRank is also $O(n^2)$ and theoretically, it is scalable to very large amounts of data.

3.4 Ranking Computation

The invariant probability vector represents the ranking values for all the entities in the knowledge base, and can be obtained with the power iteration method using Equation (2) (We shall refer the ranking scores to as "RareRank" scores). The initial values in the rank vector π_0 can be set to all $1/N$s, alternatively, one of the elements in the rank vector is set to 1, and all others to 0. After a number of iterations, probability values in the rank vector start to converge to the invariant distribution, and are irrelevant to the initial values.

3.5 Ranking Entities in RareRank

Semantic search generalises conventional retrieval systems from retrieving and ranking of documents (e.g., Web pages and scholarly articles) to entities (e.g., documents, person, institute, etc). By using technologies introduced in the semantic Web research (e.g., ontologies), the RareRank approach harmonises different semantically related entities and provides a solution for generating meaningful rankings.

Besides producing document ranking that integrates quality[9] and relevance, RareRank is also able to produce rankings for other entities presented in the knowledge base (e.g., publications, researchers, journals and conferences), especially, rankings of entities reinforce each other in an iterative procedure. To some extend, it also generalises some of the existing applications such as expert finding (using language models [29], probabilistic models [30]), and journal ranking using the Impact Factor [31]. It provides an alternative approach for these existing applications by integrating the different tasks in a coherent framework.

4 Experiment

Experiments have been conducted to generate ranking results for different types of entities in the knowledge base using the RareRank. We first present the experimental settings including the dataset preparation and configuration of parameters used in the algorithm. Then

[9] Citation analysis has been widely used as the primary method for assessing quality of published works

we show the computation of the probability transition matrix and ranking vector.

4.1 DataSet

We used the IRIS2 publication ontology and knowledge base ACM-SW [28] to evaluate the RareRank algorithm in our experiment. A topic ontology learned using an ontology learning method [25] was implanted into the knowledge base. The knowledge base was then saved in a repository and each entity was assigned a unique identifier. Statistics of the data contained in the knowledge base is shown in Table 1.

TABLE 1
Statistics of the knowledge base for entity ranking

Name	Number
Number of nodes in K	6,858
publication nodes	4,017
topic nodes	77
author nodes	1,830
publisher nodes	934
Number of relations in K	41,355
cite	4,269
hasTopic/isTopicOf	5220
isWittenBy/write	23,698
broader/narrower/related	442
publish/publishedIn	7,726

4.2 Default Parameter Setting

Weights of the predicate links defined the ontology schema graph are customisable (manually set) parameters in RareRank[10]. They essentially reflect the users' search preferences and the semantics of a domain [5]. A typical setting of predicate weights is shown in Figure 1. In the default setting, the weight of the link "iris2:hasTopic" is set to 0.55, and weight of "iris2:cite" is set to 0.3. This reflects that relevancy of publications with the topics has been emphasised and the effect of citations has been degraded. If citation (quality) is more preferable over relevancy, the weight can be set to higher values (If citation link is set to 1 and other links are set 0, then RareRank restores the original PageRank). From the publication node, users might also navigate through the links "iris2:publishedIn" and "iris2:isWrittenBy" to the publisher and author nodes with different transition probabilities. The transition modelling also reflects the occasional behaviour of users who search published works by browsing conference proceedings, journal issues and authors' publication lists. Weights for these two kinds of links are set much lower than others in this exemplar scenario. If a user wants to navigate from one publication to another with a similar topic, he traverses to the topic node through the outgoing link "iris2:hasTopic" and then traverses to another publications through the link "iris2:isTopicOf". The links "skos:broader", "skos:narrower", and "skos:related" are used to model a topic map in users' mind when they are engaged in research activities.

The weights of the predicates can also be estimated automatically using more sophisticated approaches, such as monitoring community of users' search activities by collecting user clickthrough data. After sufficiently long time period, probability transition values can then be computed from the collected data which would reflect real search patterns of an "average" user. We skip detailed discussion on this issue since it is not the focus of this paper.

4.3 Relating Topics and Documents

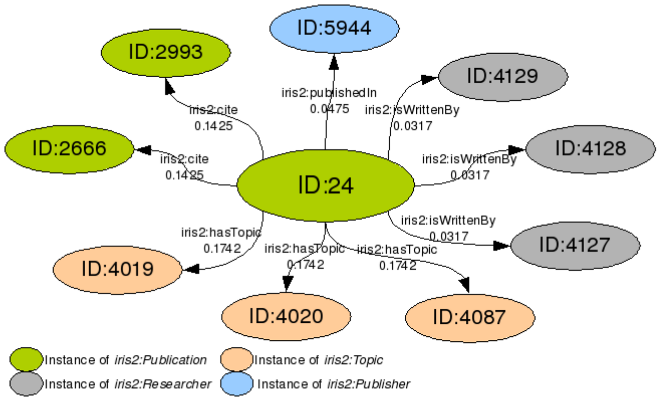
Relationships between the topics and documents were modelled using the predicate "iris2:hasTopic" and its inverse predicate "iris2:isTopicOf". Topics and documents were represented as low dimensional vectors by using Latent Dirichlet Allocation (LDA) [10], [11] as a dimension reduction technique. We then calculated the weights of predicates between topics and documents using their LDA representations. For each topic, similarity values between itself and all the documents are calculated using the Cosine similarity measure [32] (the similarity also can be calculated using the divergence measure such as Jason-Shannon Divergence [33]). Documents with similarity values greater than a threshold were selected and linked to the corresponding topics using the "iris2:hasTopic" relationships. Those documents with lower similarity measure were considered as irrelevant in terms of content.

4.4 Computing Probability Transition Matrix

With the parameter settings and the algorithm for computing transition probability vectors, it was straightforward to compute the transition probability matrix. An example of the transition probabilities of an entity to other entities are demonstrated in Figure 2.

Values of the transition probabilities for different types of links are shown in Figure 2(a) and denotations (URIs) of the node Ids are shown in Figure 2(b). The node $ID24$ links to two other publication nodes via "iris2:cite", three topic nodes via "iris2:hasTopic", three author nodes via "iris2:isWrittenBy", and one publisher node through "iris2:publishedIn". The sum of the weights of all predicates of the same type equals to the one predefined in the schema ontology (note that the values in the figure has been multiplied by the damping factor 0.95). For example, in the publication ontology, weight of the predicate "iris2:cite" is 0.3; in Figure 2(a), there are two instances of the predicate, the value is: $2 * 0.1425/0.95 = 0.3$. Sum of all the transition probabilities is equal to 0.95. Another 0.05 is

[10] Due to the limited number of links in the schema graph, the amount of manual setting work is trivial.



(a) Transition probabilities

ID:24	- *iris2:acmp_http://doi.acm.org/10.1145/1096601.1096624*
ID:2666	- *iris2:acmp_http://dx.doi.org/10.1162/0148926042728458*
ID:2993	- *iris2:acmp_http://dx.doi.org/10.1162/0148926054094369*
ID:4019	- *iris2:classification*
ID:4020	- *iris2:clustering*
ID:4087	- *iris2:visualization*
ID:4127	- *iris2:Andruid_Kerne*
ID:4128	- *iris2:Eunyee_Koh*
ID:4129	- *iris2:J_Michael_Mistrot*
ID:5944	- *iris2:DocEng_05_the_2005_ACM_symposium_on_Document_engineering*

(b) Node denotation

Fig. 2. An example of transition probabilities of nodes in knowledge base

the probability for teleport operation[11]. Each of the nodes has a probability of 0.05/6858=7.290755322251386E-6 to initiate the teleport operation.

Due to the large amount of memory needed in constructing the transition probability matrix, it is decomposed and saved into two matrices (stored as files): teleport matrix (each row has the same value and we represent it as a vector) and jumping matrix (a sparse matrix).

4.5 Ranking Vector

The ranking vector was generated by first loading the transition probability matrix from the two matrix files, and then applying the power iteration method. In our experiment, after about 20 iterations, the ranking vector started to converge to its invariant distribution, regardless of the initial values.

5 Evaluation

We used human judgment of relevance to evaluate the produced rankings. 60 queries were prepared for the evaluation and the retrieved documents were ranked using the RareRank scores which represents their global importance. We adopted two strategies for retrieving the documents: one utilised a text-based search engine built on Lucene[12], and another computed similarity values between the query and documents using their low dimensional representations based on the LDA model.

In this section, we first explain the methods used for assessing performance of the ranking algorithms. Then we present the evaluation results for the ranked entities, emphasising publications. We have also implemented the ObjectRank [5] and the original PageRank algorithm [1], [2], and compared the experimental results generated using RareRank with those generated using ObjectRank and the original PageRank.

5.1 Evaluation Methods

General Information Retrieval measures for assessing performance of text retrieval systems, such as recall, precision and $F1$ [32], [16] are not sufficient to assess the performance of ranking algorithms. The first measure we considered is the Precision at n, or $P@n$, defined as the precision at the cut-off value n. The measure reflects the actual measured system performance as a user might see it [34].

Another measure we considered is the Normalised Discounted Cumulative Gain ($NDCG_n$) [35]. The measure is designed based on the intuition that since all documents are not of equal relevance to users, highly relevant documents should be identified and ranked first for presentation to the users [35]. It adopts graded relevance assessments, as opposed to traditional evaluation methods, such as recall and precision which are based on binary relevance assessments, and thus credits IR methods for their ability to retrieve highly relevant documents quickly. The $NDCG_n$ is calculated using Equation (11).

$$NDCG_n = \frac{DCG_n}{IDCG_n} \tag{11}$$

where DCG_n is the Discounted Cumulative Gain, and $IDCG_n$ is the Ideal Discounted Cumulative Gain, which is calculated as the discounted cumulative gain of an ideal ranking. DCG_n is calculated using Equation (12).

$$DCG_n = \sum_{i=1}^{n} \frac{2^{label(i)} - 1}{log_b(1+i)} \tag{12}$$

where $label(i)$ is the gain value associated with the label of the document at the ith position of the ranked list. The discounting factor b allows modeling user impatience (a small value of b, e.g., $b = 2$) and persistence ($b = 10$)[13]. Empirical studies on $IDCG_n$ [35] has shown that $IDCG_n$ conveys more credit to systems with high precision at top ranks than other evaluation measures. In our evaluation, we set $b = 2$, and used a graded relevance judgment, with $label(i) = 2$ corresponding to "highly relevant", 1 corresponding to "moderately relevant", and 0 corresponding to "irrelevant".

[11] It equals to the base teleport probability. The "schema imbalance" and "link Zero-instantiation" teleport probabilities are both 0.

[12] http://lucene.apache.org/

[13] Smaller values of b cause greater discounting of documents retrieved at lower ranks.

5.2 Evaluation Results

The RareRank scores represent global importance of entities of different types in the knowledge base. In the following we report the experimental results on document entities using the evaluation measures defined earlier. The results are compared with the ObjectRank and PageRank algorithms. Furthermore, we present and discuss the RareRank scores for author and publisher entities.

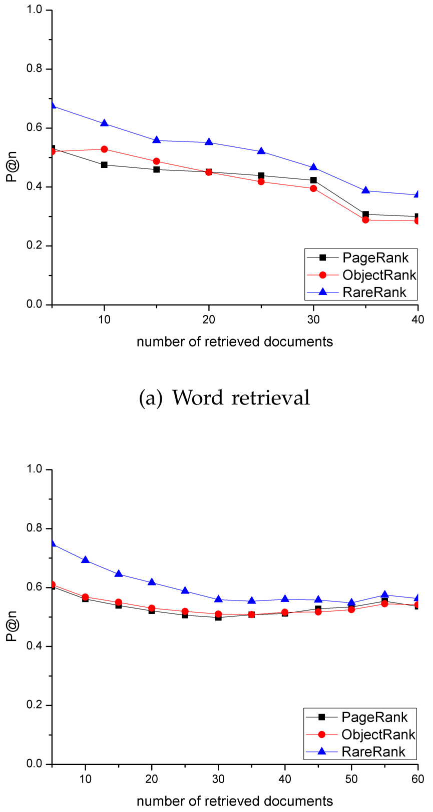
5.2.1 Precision Measures

We prepared 60 popular search terms related to the semantic Web, Information Retrieval, and machine learning, and retrieved documents using two strategies: the first strategy retrieved documents based on a content-based search engine using Lucene (referred to as word retrieval), and the second strategy retrieved documents by selecting those whose similarity measures with the queries are greater than a threshold (referred to as topic retrieval), then expanding the initial document set using links in the knowledge base graph. For each strategy, documents were ranked using the RareRank and PageRank scores. For word retrieval we only evaluated the 40 top ranked documents, and for topic retrieval, we evaluated 60 top ranked documents using $P@n$ and $NDCG$ measures (Some of the word retrieval generated less than 40 results). We also conducted experiments using ObjectRank and PageRank using the same dataset and similar parameter settings. Note in both ObjectRank and PageRank, the idea of using terminological topic ontology in combination with knowledge base for the purpose of ranking was not introduced. $P@n$ and $NDCG$ values generated using the three algorithms are shown in Figure 3 and 4, respectively. In the experiments, the word retrieval method generally produced low precision compared to the topic retrieval method, however, it is useful when topic retrieval does not return any results.

5.2.2 Comparison Study

The $P@n$ and $NDCG_n$ values computed under different search strategies were averaged across all the queries. The averaged $P@n$ values of RareRank, ObjectRank, and PageRank using two retrieval strategies are shown in Figure 3. Using word retrieval, 27 out of 60 queries returned more than 40 documents; while using the topic retrieval, 48 out of 60 queries returned more than 40 documents and 14 out of 60 queries returned more then 60 documents.

Figure 3(a) and 3(b) show that at all document cutoff levels, $P@n$ values using RareRank are higher than those of ObjectRank and PageRank. It is unexpected that performance of the original PageRank is comparable to ObjectRank in terms of $P@n$ measure. Similar pattern can be observed in terms of $NDCG$ measure as shown in Figure 4(a) and 4(b). A possible explanation is that in our implementation of the ObjectRank we only considered the "global ObjectRank" [5]. The main reason that we did not make use of "keyword-specific ObjectRank" [5] is that matching individual terms in queries with words in title of publications only is not effective because breaking down the key-phrases into individual terms destroys their intended meanings, especially for searching in the domain of scientific research. Furthermore, computation for the "keyword-specific" ObjectRank is very expensive. Figure 3(b) also shows that $P@n$ values of RareRank approach those of ObjectRank and PageRank at document cutoff levels from 45 to 60.



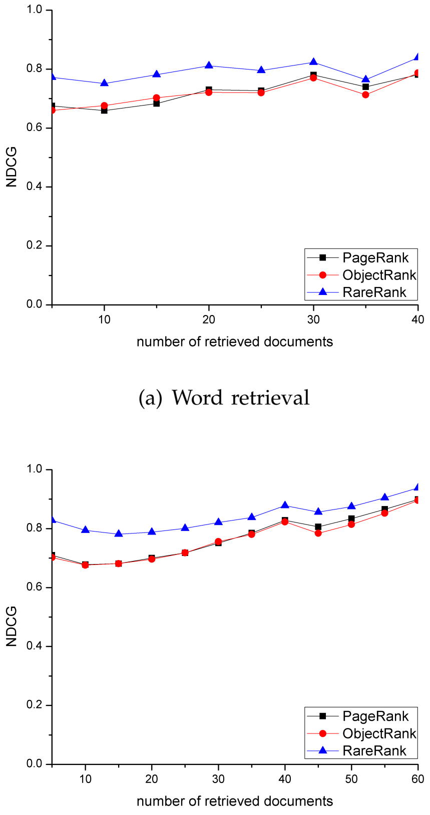
(a) Word retrieval

(b) Topic retrieval

Fig. 3. Averaged $P@n$ values using RareRank and PageRank

The averaged $NDCG_n$ values of RareRank, ObjectRank, and PageRank using two retrieval strategies are shown in Figure 4. The $NDCG_n$ values of RareRank are higher than those of PageRank at all document cutoff levels. In Figure 3 and 4, the tails of the $P@n$ and $NDCG_n$ curves at document cutoff level of 35 and 45 demonstrate some strange behaviour, i.e., notable falls. Examining query results, we found that some of the queries produce less than 35 and 45 documents using the word and topic retrieval respectively. The resulting averaged $P@n$ and $NDCG_n$ values thus demonstrate "inconsistency" at these two cutoff points.

To determine whether the observed differences between the three ranking approaches are statistically significant, we performed statistical significance tests using



(a) Word retrieval

(b) Topic retrieval

Fig. 4. Averaged $NDCG_n$ values using RareRank and PageRank

TABLE 2
Statistical significance tests of RareRank, ObjectRank, and PageRank using paired student T-test at significance level of 0.05

Comparison	**t values**					
	RareRank vs ObjectRank		RareRank vs PageRank		ObjectRank vs PageRankRank	
	$P@n$	$NDCG$	$P@n$	$NDCG$	$P@n$	$NDCG$
Value	5.6732	9.9371	5.2696	8.2505	1.3418	2.7434

the paired T-test. The results calculated using both averaged $P@n$ and $NDCG$ values are reported in Table 2. By conventional criteria, differences between RareRank and ObjectRank and PageRank are considered as statistically significant at the significance level of 0.05 (the differences are also significant at level of 0.001), no matter $P@n$ ($t = 5.6732$ and 5.2696) or $NDCG$ ($t = 9.9371$ and 8.2505) values are used. At level of 0.05, difference between ObjectRank and PageRank is not significant when $P@n$ values are used for the paired t-test ($t = 1.3418$); while the difference is significant when $NDCG$ values are used ($t = 2.7434$). This is due to the cumulative nature of the $NDCG$ calculation. However, at the level of 0.001, the difference is not significant any more when calculated using $NDCG$ values. The statistical significance test demonstrates the superior performance of RareRank for ranking over ObjectRank and PageRank in this comparison study.

5.2.3 Researcher and Publisher Ranking

Beside publication ranking, the "Rational Research" model is able to produce ranking for other entities, such as researcher and publisher. The objective in this paper is not to provide a complete enumeration of prominent researchers and publishers in the semantic Web research area, but to demonstrate the effectiveness of the proposed model for ranking entities in semantic search applications.

Table 3 illustrates the top 10 researchers in the semantic Web area ranked by RareRank. Note that the rankings are completely dependent on the underlying dataset used in our experiment (which is neither complete nor error-free). As shown in the table, there is no obvious correlation between the ranking of researchers and the number of publications they have in the dataset. Intuitively, the ranking of researchers is affected by the rankings of its surrounding entities, e.g., publications. The researcher "Sheila A. McIlraith" (with 7 publications in the dataset) is ranked highly because one of her publications "Semantic Web Services" has been cited 121 times in our dataset (ACM Digital Library record, Google Scholar reports 1091 citations), which is the most significant citations count compared to others. This matches the intuition that ranking of entities reinforces each other in the RareRank algorithm.

TABLE 3
Ranking of researchers

Ranking	Name	Num Pub	RareRank	PageRank
1	Sheila A. McIlraith	7	2.850	7.762
2	Steffen Staab	29	2.318	17.643
3	Tran Cao Son	2	2.282	4.175
4	Hai Zhuge	19	2.204	5.728
5	James Hendler	12	2.104	15.640
6	Ian Horrocks	25	2.017	19.376
7	Erhard Rahm	7	1.583	14.058
8	Dieter Fensel	27	1.570	16.709
9	Amit Sheth	22	1.562	7.023
10	Alexander Maedche	13	1.556	11.630
11	Philip A. Bernstein	6	1.493	12.906
12	Stefan Decker	20	1.420	15.634
13	Natalya F. Noy	11	1.399	9.454
14	Munindar P. Singh	11	1.398	3.732
15	Mark A. Musen	12	1.352	10.167
16	Enrico Motta	24	1.348	8.975
17	Wolfgang Nejdl	17	1.328	8.761
18	Katia Sycara	17	1.319	8.065
19	Alon Halevy	16	1.315	8.631
20	Anupam Joshi	22	1.311	10120

Although we cannot judge which method for researcher ranking is more preferable, we are confident with the results generated using RareRank: it correctly produces a list of high-profile researchers, and the top ranked researchers are indeed prominent people in the domain of study.

Table 4 and 5 shows a number of prominent journals and conferences in which semantic Web researchers frequently publish their research results (IF2008 is the 2008 journal impact factor[14]).

TABLE 4
Ranking of journals

Ranking	Journal name	RareRank	ObjectRank	IF2008
1	IEEE Intelligent Systems	13.810	75.535	2.3
2	Data/Knowledge Engineering Elsevier	4.901	16.783	1.5
3	IEEE Internet Computing	4.740	23.180	2.3
4	Web of Semantics	3.962	13.634	3.0
5	Communications of the ACM	3.956	28.855	2.6
6	The Knowledge Engineering Review	3.878	20.762	1.6
7	The Very Large DataBase Journal	3.719	24.060	6.8
8	Int. J. of Human Computer Studies	3.333	11.599	1.8
9	Future Generation Computer System	2.694	5.766	1.5
10	BT Technology Journal	2.218	7.493	0.4

Table 4 shows that RareRank produces slightly better predictive values than ObjectRank in terms of journal ranking (using IF2008 as a baseline). The comparison result shows that RareRank does demonstrate its capability to predicting journal ranking with reasonable correctness, even though we cannot conclude that RareRank has comparable predictive power with IF2008 or more predictive power than ObjectRank, due to the limited range and size of the dataset (compared to the one used by IF2008).

TABLE 5
Ranking of conferences

Ranking	Conference name	RareRank	ObjectRank
1	World Wide Web 03	5.106	20.267
2	2006 IEEE/WIC/ACM on Web Intelligence	4.961	15.149
3	World Wide Web 04	4.369	16.666
4	World Wide Web 06	4.225	15.151
5	World Wide Web 02	3.556	17.874
6	2007 IEEE/WIC/ACM on Web Intelligence	3.510	17.222
7	2004 IEEE/WIC/ACM on Web Intelligence	2.929	14.443
8	World Wide Web 05	2.894	7.454
9	World Wide Web 07	2.884	9.897
10	1st International Semantic Web Conference	2.357	19849

In Table 5 conference "WWW 03" was ranked at the first place. A reasonable explanation is that at that time research on semantic Web has attracted attention of many researchers, and many papers related to the semantic Web have been published at that conference. Moreover, some of the papers published in the semantic Web track in that year such as "Semantic Search" by Guha *et al*, and "Agent-based Semantic Web Services" by Gibbins *et al*, have been cited many times over the past few years (Google Scholar reports 225 and 112 citations for the two papers respectively).

Evaluation of the rankings of researchers and publishers is especially problematic due to the subjective nature of the task, availability of large number of influencing factors, and difficulty in finding optimal parameter combinations. To our knowledge, currently there are no standard evaluation methods for the task and most of the existing works evaluate the rankings human judgement of relevance. The problem of publisher ranking, in particular, journal ranking has been studied in large number of works. One of the most authoritative journal ranking methods is based on the impact factor [31], which is computed using statistics on citations to a specific journal. However, ranking based on the impact factor also has some limitations: there is no ranking for conferences, and computation of the impact factor is a sluggish process. On the contrary, the "Rational Research" model generalises the notation of ranking in different contexts (e.g., such as expert finding and journal ranking), and harmonises the tasks of ranking different types of objects.

5.3 Remarks on Retrieval Quality

Evaluation based on $P@n$ and $NDCG_n$ measures mostly reveals relevancy of the retrieved results. For research publication retrieval, quality evaluation is always a subjective and difficult procedure. Currently, the most prevalent measure for assessing quality of scholarly articles is the citation counts. However, citation count is not the only factor that determines the quality.

In scientific research, establishment of a citation link is based on human cognitive process such as judgement of novelty, aknowledgement for original contribution, or even criticism. In addition to the publication pipeline delay and time spent to read the papers, cumulation of citation counts is often a prolonged process. In today's competitive research environment, a publication with perceived high quality (many citations) may not be very relevant to the state-of-the-art after several years. Therefore, the desirable publications should have characteristics of both quality and relevance. Intuitively, the RareRank algorithm favors those documents with reasonable number of citations, and strongly relevant content related to user's query. Our experimental results indeed reveal such intuition: the top ranked results for a specific query are those having balance between relevance and quality. Another characteristic of RareRank is that even a newly written document could obtain a high rank value. Consequently, it is able to promote the presence and dissemination of newly-written documents that have not been aware of and not been cited by many other authors.

6 Conclusion and Future Work

Today's search engines rely on ranking algorithms to select quality and relevant results from large document repositories in responding to user queries. Many ranking algorithms, in particular, link analysis, have been developed during the past decades, and have been proved as effective and scalable means for ranking documents in modern retrieval systems. Semantic search generalises traditional IR from pure document to entities search and

[14] http://abhayjere.com/Documents/Impact factor 2008_PDF.pdf

retrieval, and poses an additional challenge on the capability of retrieval systems: to retrieve and rank entities of various types. In this paper, We present the idea of the "Rational Research" model and develop the RareRank algorithm to address the challenge (in the context of scientific research). In "Rational Research", a terminological topic ontology is added into the knowledge base to simulate a research environment, and the relationships between various entities simulate the behaviour of a "rational researcher" who undertakes his research activities. Computation of the RareRank scores is based on a set of rules for computing the transition probability matrix and is guaranteed to converge to an invariant distribution. Experimental study has shown that in terms of two ranking measures, Precision at n, and Normalised Discounted Cumulative Gain, RareRank outperformed ObjectRank and the original PageRank algorithms. Nevertheless, the size of the dataset used in our current study is limited. We believe that using existing large datasets developed by the research community would add more credibility to our work. Our future work would involve investigating RareRank's applicability and performance on larger data collections.

Acknowledgments

The authors would like to thank all the reviewers for their insightful comments and valuable suggestions. The authors also would like to thank the staffs and researchers of the Intelligent Decision Support Systems research group at the University of Nottingham Malaysia Campus, and the Centre for Communication Systems Research at the University of Surrey, UK for their help on evaluating the experimental results.

References

[1] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," Stanford Digital Library Technologies Project, Tech. Rep., 1998.

[2] A. Langville and C. Meyer, "Deeper inside pagerank," *Internet Mathematics*, vol. 1, no. 3, pp. 335–380, 2004.

[3] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," in *SODA*, 1998, pp. 668–677.

[4] S. Lawrence, C. L. Giles, and K. Bollacker, "Digital libraries and Autonomous Citation Indexing," *IEEE Computer*, vol. 32, no. 6, pp. 67–71, 1999.

[5] A. Balmin and V. Hristidis, "Objectrank: Authority-based keyword search in databases," in *In VLDB*, 2004, pp. 564–575.

[6] H. Hwang, V. Hristidis, and Y. Papakonstantinou, "Objectrank: a system for authority-based search on databases," in *SIGMOD Conference*, 2006, pp. 796–798.

[7] S. E. Robertson, S. Walker, and M. Hancock-Beaulieu, "Okapi at trec-7: Automatic ad hoc, filtering, vlc and interactive," in *TREC*, 1998, pp. 199–210.

[8] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman, "Indexing by latent semantic analysis," *JASIS*, vol. 41, no. 6, pp. 391–407, 1990.

[9] T. Hofmann, "Probabilistic latent semantic analysis," in *UAI*, 1999, pp. 289–296.

[10] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[11] M. Steyvers and T. Griffiths, "Probabilistic topic models," in *Latent Semantic Analysis: A Road to Meaning*, T. Landauer, D. Mcnamara, S. Dennis, and W. Kintsch, Eds. Laurence Erlbaum, 2005.

[12] S. Wasserman and K. Faust, *Social network analysis: methods and applications*, 1st ed. Cambridge: Cambridge Univ. Press, 1997.

[13] H. Small, "Co-citation in the scientific literature: A new measure of the relationship between two documents," *Journal of the American Society for Information Science*, vol. 24, no. 4, pp. 265–269, 1973.

[14] E. Garfield, "From computational linguistics to algorithmic historiography," 2001, paper presented at the Symposium in Honor of Casimir Borkowski at the University of Pittsburgh School of Information Sciences.

[15] H. Nanba and M. Okumura, "Automatic detection of survey articles," in *Research and Advanced Technology for Digital Libraries*. Springer, 2005, pp. 391–401.

[16] C. D. Manning, P. Raghavan, and H. Schôtze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[17] C. Andrieu, N. de Freitas, A. Doucet, and M. I. Jordan, "An introduction to mcmc for machine learning," *Machine Learning*, vol. 50, no. 1-2, pp. 5–43, 2003.

[18] T. Haveliwala, "Topic-sensitive PageRank: A context-sensitive ranking algorithm for web search," *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 4, pp. 784–796, 2003.

[19] M. Richardson and P. Domingos, "The Intelligent Surfer: Probabilistic Combination of Link and Content Information in PageRank," in *Advances in Neural Information Processing Systems 14*. MIT Press, 2002.

[20] L. Ding, T. Finin, A. Joshi, R. Pan, R. S. Cost, Y. Peng, P. Reddivari, V. Doshi, and J. Sachs, "Swoogle: a search and metadata engine for the semantic web," in *CIKM '04*, New York, NY, USA, 2004, pp. 652–659.

[21] L. Ding, R. Pan, T. Finin, A. Joshi, Y. Peng, and P. Kolari, "Finding and ranking knowledge on the semantic web," in *Proceedings of the 4th International Semantic Web Conference*, ser. LNCS 3729. Springer, November 2005, pp. 156–170.

[22] A. Hogan, A. Harth, and S. Decker, "Reconrank: A scalable ranking method for semantic web with context," in *Proceedings of SSWS2006*, 2006.

[23] A. Harth, A. Hogan, R. Delbru, J. Umbrich, S. ORiain, and S. Decker, "Swse: Answers before links!" in *Proceedings of Semantic Web Challenge*, 2007.

[24] T. Lukasiewicz and J. Schellhase, "Variable-strength conditional preferences for ranking objects in ontologies," *J. Web Sem.*, vol. 5, no. 3, pp. 180–194, 2007.

[25] W. Wang, P. M. Barnaghi, and A. Bargiela, "Probabilistic topic models for learning terminological ontologies," *IEEE Transactions on Knowledge and Data Engineering*, vol. 99, no. PrePrints, 2009.

[26] M. Bender, T. Crecelius, M. Kacimi, S. Michel, T. Neumann, J. X. Parreira, R. Schenkel, and G. Weikum, "Exploiting social relations for query expansion and result ranking." in *ICDE Workshops*. IEEE Computer Society, 2008, pp. 501–506.

[27] M. J. Bates, "The design of browsing and berrypicking techniques for the online search interface," *Online Review*, vol. 13, no. 5, pp. 407–424, 1989.

[28] W. Wei, "Semantic search: Bringing semantic web technologies to information retrieval," Ph.D. dissertation, School of Computer Science, The University of Nottingham, 2009.

[29] K. Balog, L. Azzopardi, and M. de Rijke, "Formal models for expert finding in enterprise corpora," in *SIGIR*, 2006, pp. 43–50.

[30] H. Fang and C. Zhai, "Probabilistic models for expert finding," *LECTURE NOTES IN COMPUTER SCIENCE*, no. 4425, pp. 418–430, 2007.

[31] E. Garfield, "Journal impact factor: a brief review." *Canadian Medical Association journal (CMAJ)*, vol. 161, pp. 979–980, 1999.

[32] R. A. Baeza-Yates and B. A. Ribeiro-Neto, *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999.

[33] J. Lin, "Divergence measures based on the shannon entropy," *IEEE Transactions on Information Theory*, vol. 37, no. 1, p. 145, 1991.

[34] E. Voorhees, "Overview of the trec 2006: Common evaluation measures," in *Proceeding of The Fifteenth Text REtrieval Conference*, 2006.

[35] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of ir techniques," *ACM Trans. Inf. Syst.*, vol. 20, no. 4, pp. 422–446, 2002.