MedRank: Discovering Influential Medical Treatments from Literature by Information Network Analysis

Ling Chen[1,2] **Xue Li**[1,3] **Jiawei Han**[4]

[1] School of Information Technology and Electrical Engineering
University of Queensland, Australia
[2] Email: lchen5@uq.edu.au
[3] Email: xueli@itee.uq.edu.au

[4] Department of Computer Science
University of Illinois at Urbana Champaign
Urbana, United States
Email: hanj@uiuc.edu

Abstract

Medical literature has been an important information source for clinical professionals. As the body of medical literature expands rapidly, keeping this knowledge up-to-date becomes a challenge for medical professionals. One question is that for a given disease how can we find the most influential treatments currently available from online medical publications? In this paper we propose MedRank, a new network-based algorithm that ranks heterogeneous objects in a medical information network. The network is extracted from MEDLINE, a large collection of semi-structured medical literature. Different types of objects such as journal articles, pathological symptoms, diseases, clinical trials, treatments, authors, and journals are linked together through their relationships. The experimental results are compared with the expert rankings collected from doctors and two baseline methods, namely degree centrality and NetClus. The evaluation shows that our algorithm is effective and efficient. The success of categorized entity ranking in medical literature domain suggests a new methodology and a potential success in ranking semi-structured data in other domains.

1 Introduction

The vast body of medical literature grows rapidly every year. Taking MEDLINE the premier bibliographic database of the world's largest medical library, supported by the U.S. National Library of Medicine (NLM), as an example, there are medical journal articles (10,390,997), clinical trials (1,011,711), references to diseases (587,012), and publication of 5,400 international journals in MEDLINE 2010. About 700,000 new records were added into MEDLINE in 2011. The MEDLINE 2012 baseline now has more than 20 million records.[1]

The direct implication of this trend is that it is becoming more and more difficult for doctors to keep their medical knowledge up-to-date by processing information manually. They are facing a challenge of accessing relevant information for evidence-based decision support. Many studies have found improvements for search functionality in existing medical databases by using information retrieval techniques (Hliaoutakis et al. 2009, Luo 2009, Luo & Tang 2008, Luo et al. 2008). However, the core problem lies not in retrieving best-matched records, but in knowledge discovery.

In this research, we approach the literature-based knowledge discovery problem by tackling one of its example problems, that is, finding the most influential treatments for a given disease. A treatment is "influential" if it is mentioned by many *good* articles and published with clinical trials that have positive results. A *good* article is one that is written by reputed author(s) and published in a *good* journal. A *good* medical journal is one that has a high impact on research. In our research we use the MEDLINE database to construct a medical *information network* (Easley & Kleinberg 2010, Sun, Yu & Han 2009) that is abstracted as a graph with referential relationships amongst different types of objects extracted from MEDLINE. In order to find the *most* influential treatments, say the top-10, all associated objects have to be ranked. Thus, our problem becomes a ranking problem: Given a disease name, how could we rank the most influential treatments?

Existing ranking methods can be classified into three categories, namely *preference-based*, *similarity-based* and *network-based*.

- **Preference-based ranking** has been studied for a long time since 1904 (Spearman 1904). Preference-based rankings always reflect subjective views on objects by humans. In this ranking process, preferences of objects are collected and then a model will be derived (Ceci et al. 2010).
- In **similarity-based ranking**, objects are expressed as vectors of their attributes. For example in Collaborative Filtering (Shardanand & Maes 1995), Cosine or Euclidean distance functions can be used to rank objects. Similarity-based rankings are widely used in recommendation systems (Adomavicius & Tuzhilin 2005).
- **Network-based ranking** was first referred to as PageRank (Page et al. 1999) and HITS (Kleinberg 1999) algorithms. In this kind of ranking, a graph structure is used. An iterative process is applied to propagate *good* properties of objects through links. Network-based ranking is sometimes referred to as the link-analysis based rank-

[1] http://www.nlm.nih.gov/bsd/licensee/2012_stats/baseline_doc.html

ing. An insightful introduction can be found in (Borodin et al. 2005).

In practice, there are often mixed models, for example, the current search methods used by Google. In this paper, a new network-based ranking algorithm called MedRank is proposed to find the most influential treatments for a given disease. To the best of our knowledge, this is the first work that introduces information network analysis techniques for ranking objects in a medical domain. A multi-dimensional medical information network is constructed based on the categorized entities identified in medical literature. The proposed algorithm has been evaluated against expert rankings collected from doctors and the baseline methods, i.e., degree centrality and NetClus. The evaluation shows that MedRank outperforms both baseline methods.

The rest of the paper is organized as follows: Section 2 discusses related work, followed by Section 3 that gives the problem formalization. Section 4 presents the proposed MedRank algorithm. Section 5 describes the experiments, survey analysis and evaluation, followed by discussions in Section 6. Finally, Section 7 presents the conclusions.

2 Related Work

2.1 Existing Literature-based Systems

Information Retrieval techniques have been applied to search for medical information in repositories such as PubMed[2] and UpToDate.[3] For example, Ratprasartporn et al. (2009) proposed a content-based method for digital literature collection search with experiments on PubMed. Chen et al. (2011) introduced a passage retrieval method for MEDLINE articles. Luo et al. (2008) and Luo (2009) have proposed Web search engines to find medical information over the Web. In medical literature-based knowledge discovery, MEDLINE has been used as a main source to discover relationships between the medical concepts appeared in medical journal articles (Petric et al. 2009, Yetisgen-Yildiz & Pratt 2006). All of these focus on filtering out the irrelevant information and assisting users to find out the medical articles relevant to the given medical terms.

Most clinical decision support systems are built based on the past clinical records and the analytical reasoning on the causal relationships established among the symptoms, patient demographic data, diseases, treatments, etc. Many researchers developed graph-based models, such as Bayesian Network and Artificial Neural Network, for practical systems (Berner 2007). As an exception, Zhao & Weng (2011) proposed a diagnostic system based on both PubMed and Electronic Health Records (EHRs) for predicting pancreatic cancer using Bayesian Network. Munteanu et al. (2009) introduced a star graph model for cancer-related protein classification. But it is not literature-based nor applicable to heterogeneous objects and their target is classification, not ranking.

2.2 Existing Network-based Ranking Methods

Network-based ranking was first referred to as the PageRank (Page et al. 1999) and HITS (Kleinberg 1999) algorithms. A directed graph of hyperlinked Web pages on the WWW is used. The idea behind is essentially the *eigenvector centrality* that has been long studied in Social Network Analysis (Newman 2010). Unlike the *degree centrality* that considers all neighbouring nodes equally important, the eigenvector centrality finds those "center" or important nodes such that their neighbours are themselves important. The key idea of PageRank is the *rank propagation through links*, i.e., ranks are propagated from one Web page to another through the hyperlinks. The original PageRank model is often explained as a Markov chain with a transition probability matrix. The PageRank vector is iteratively updated with the matrix until it converges to a limiting distribution. The convergence is guaranteed as the transition matrix has been shown to be irreducible (i.e., strongly connected) and aperiodic (i.e., non-bipartite) (Langville & Meyer 2004). HITS algorithm, on the other hand, aims to find authoritative pages based on a user supplied query. It considers not only the authoritative pages as PageRank does, but also the *hub* pages that have links to multiple relevant authoritative pages.

PopRank (Nie et al. 2005) extends the PageRank model from the Web page level to the Web object level and from ranking homogeneous objects to heterogeneous ones. Web objects may belong to different types, such as article or people, and be related to each other in different ways, such as cited-by, written-by, etc. As the importance of different types may differ, PopRank automatically assigns an optimized weight to every type of relationships, called the *popularity propagation factor* (PPF).

Sun *et al.* extend the ranking mechanism of PopRank from the Web objects to a network of heterogeneous objects extracted from DBLP, a bibliographic database in computer science. Unlike the previous work, an undirected graph is used. Their first work is RankClus (Sun, Han, Zhao, Yin, Cheng & Wu 2009), a ranking-based clustering algorithm that ranks bi-type objects in its own type within clusters. NetClus (Sun, Yu & Han 2009) is proposed in their later work to handle multi-type objects in a special kind of network called the *Star Network*. It is characterized by the way different types of objects are connected in a star-like shape.

To the best of our knowledge, the proposed MedRank method is the first work that introduces the network-based ranking approach to the medical domain. PageRank and PopRank are not applicable to our ranking problem, because PageRank is designed for one type only, i.e., Web page, and both of them are directly applicable only to directed graphs. MedRank's main difference from RankClus and NetClus is that it is based on the available category labels and no clustering mechanism is involved.

3 Problem Formalization

In this section, we define the problem of ranking in medical information networks and introduce several related concepts and necessary notations.

Definition 1 ***Heterogeneous Information Network***
Given a set $\Gamma = \bigcup_{t=1}^{T} X_t$ *of* T *types, where* X_t *is the set of objects belonging to the* t^{th} *object type, a graph* $G = \langle V, E \rangle$ *is called an information network if* $\Gamma = V$ *and* E *is a binary relation on* V*. A heterogeneous information network* N *is an information network with* $T \geq 2$*.*

Definition 2 ***Medical Information Network***
Given a heterogeneous information network $G =$

[2] http://www.ncbi.nlm.nih.gov/pubmed/
[3] http://www.uptodate.com

$\langle V, E\rangle$, *it is a medical information network, if* $\forall X_t \subset V, X_t$ *is medical related and* $\exists X_i, X_j \subset V$ *such that* X_i *is the set of "Disease" type objects, and* X_j *is the set of "Treatment" type objects.*

Definition 3 ***Star Network***
Given a heterogeneous information network $G = \langle V, E\rangle$ *on* $(T+1)$ *types of objects and* $V = \bigcup_{t=0}^{T} X_t$, G *is called a star network if* $\forall e = \langle x, y\rangle \in E, x \in X_0 \wedge y \in X_t (t \neq 0)$, *or vice versa.*

Star-shaped structures are often found in information networks, for example, the bibliographic information networks. A star network is characterized by its edges only existing between one special type of objects, called the center type objects (i.e., X_0), and objects of other types, called the attribute type objects (i.e., $X_t (1 \leq t \leq T)$). This characteristic essentially forms a bipartite graph with a star schema. In this paper, we construct a medical star information network that has "Article" objects as the center type objects and "Disease", "Treatment", "Author", "Clinical Trial" and "Journal" objects as attribute type objects. The network schema is given in Fig. 1.

Definition 4 ***Disease Sub-network***
Given a medical star information network $G = \langle V, E\rangle$ *where* $\exists X_d \subset V$ *is the set of "Disease" type objects and given a disease name* $q \in X_d$, *a disease sub-network* N' *is a graph* $G' = \langle V', E'\rangle \subseteq G$ *such that* $V' = V - \{x \in X_0 | \neg\exists y(\langle x, y\rangle \in E \wedge y = q)\}$ *and* $E' = \{\langle x, y\rangle \in E | x, y \in V'\}$.

As a treatment is always a treatment against some disease, it makes more sense to rank treatments for a disease. Thus, we define a sub-network $N' = \langle V', E'\rangle$ of a given disease q such that every Article node in V' is an article about the disease q. This is done by subtracting any Article node $x \in X_0$ from V such that none of the Disease node y that x links to is q. Fig. 2 gives an example of an AIDS disease sub-network with two articles. Now, the problem can be formalized as:

Problem Definition 1 *Given a sub-network* $N' = \langle V', E'\rangle$ *of disease* q *such that* $\exists X_j \subset V'$ *is the set of treatment objects, a ranking function* R *from* X_j *to* $\mathbb{R}^+$, *and a number* K, *find a set* $X' \subset X_j$ *such that* $|X'| = K$ *and* $\forall y \in X', R(y) > R(x), \forall x \in (X_j - X')$.

So, the problem is to find the top-K highest ranked treatments for a given disease. In the following, when context is clear, we use X_t to denote the object set and its type name interchangeably.

4 ALGORITHMS

The proposed MedRank algorithm utilizes the linkage information among data objects to rank influential treatments for a given disease. These data objects are extracted from the medical literature as a medical information network. The goal is to rank every "Treatment" object based on its relationships with other types of objects. The essence of the algorithm is computing eigenvector centrality, which gives higher ranks to nodes whose neighbours are themselves ranked higher. The output is a list of top-k ranked treatments for the disease.

Two phases are involved, namely the network extraction phase and the ranking phase. Details of each phase are given in the following sub-sections.

Figure 1: A star network schema

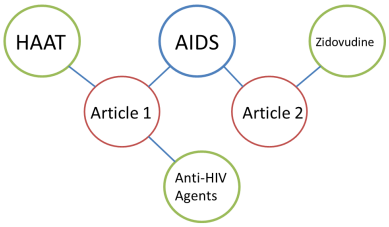


Figure 2: A disease sub-network example

Algorithm 1 NetworkExtraction

Input: L: the literature corpus; O: the ontology database; A: a set of attributes for ranking
Output: N: the extracted medical information network

1: $N \leftarrow \emptyset$; $n \leftarrow size(L)$
2: **for** $record = 1$ to n **do**
3: $\quad N \leftarrow addExtractedValues(record, A)$
4: **end for**
5: **return** N

4.1 Network Extraction

In this sub-section, we explain how a medical information network is extracted from input literature corpus using a medical ontology. The goal is to scan the corpus once and build a network for the ranking phase.

A medical literature corpus stores medical research publications up-to-date. Well-known examples are MEDLINE and Cochrane systematic reviews.[4] A medical ontology data-base provides a standardized set of medical thesaurus hierarchically structured for the classification purpose. Examples of widely used thesaurus systems include MeSH,[5] SNOMED-CT,[6] ICD-10,[7] etc.

In order to explore the relationships among objects extracted from medical literature and thus rank them accordingly, we found that articles play the role as the intermediate that connects other objects identified from them. For example, an article may be about AIDS and some treatments; by finding all the articles that discuss AIDS, we can find all the possible treatments related to AIDS in the literature. In fact, every article can be represented as a sub-graph of a star shape. Articles are also connected via their shared objects. For example, Fig. 2 shows a graph representation of two articles about "AIDS". They are connected via "AIDS" object (i.e., shared disease) and "Anti-viral Agent" object (i.e., shared treatment).

This type of network, characterized by its star-shaped schema, is called the *star network* (Sun, Yu & Han 2009). Article objects are called center type objects and the rest objects connected to them are called attribute type objects. The star network schema for

[4] http://www.cochrane.org/
[5] http://www.nlm.nih.gov/pubs/factsheets/mesh.html
[6] http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html
[7] http://www.who.int/classifications/icd/en/

our medical information network is presented in Fig. 1 and the network extraction phase is summarized as Algorithm 1.

4.2 Ranking Formulas

The goal of MedRank's ranking formula is to find the "center" or important nodes in a medical information network. PageRank computes such *eigenvector centrality*, but it is only applicable to homogeneous information networks.

To propagate ranks among multi-type objects in a star-shaped network, we adapt the Authority Ranking formula of NetClus. Formula 1 shows the $(h+1)^{th}$ iteration of ranks passing from type Z objects via center type C objects to type Y objects, as attribute type objects only have direct links to the center type C objects (see Fig. 2).

$$R_Y^{\langle h+1\rangle} \leftarrow W_{YC} D_{CZ}^{-1} W_{CZ} R_Z^{\langle h\rangle} \quad (1)$$

In Formula 1, R_Y and R_Z are rank vectors of type Y and type Z objects respectively; W_{YC} is an adjacency matrix such that if $\exists e = \langle y_i, c_j\rangle \in E, y_i \in Y, c_j \in C$, then $w_{ij} = 1$; otherwise, $w_{ij} = 0$. As an undirected graph is used $W_{YC} = W_{CY}$. W_{CZ} is defined in the same way. D_{CZ}^{-1} is a diagonal matrix with diagonal value equivalent to the row sum of W_{CZ}. Hence, $D_{CZ}^{-1} W_{CZ}$ is the row-normalized adjacency matrix of W_{CZ}.

Let A be a list of n attributes X_1, X_2, ..., X_n selected for ranking, where the target attribute of interest is X_1, and C be the center type as before. By chaining all the attributes in A (from index 1 to n back to 1) based on Formula 1, it allows ranks to propagate through all different types of objects and thus makes eigenvector centrality computation possible. The resulting matrix is presented as M in Formula 2. M is row-normalized.

$$M = (\prod_{t=1}^{n-1} W_{X_t C} D_{CX_{t+1}}^{-1} W_{CX_{t+1}}) W_{X_n C} D_{CX_1}^{-1} W_{CX_1} \quad (2)$$

As M is a $|X_1| \times |X_1|$ row-normalized matrix, it can be regarded as a transition matrix of a Markov Chain. And the update rule becomes Formula 3.

$$R_{X_1}^{\langle h+1\rangle} \leftarrow M R_{X_1}^{\langle h\rangle} \quad (3)$$

However, $R_{X_1}^{\langle h\rangle}$ will only converge to a long-run stationary vector $R_{X_1}^{\langle *\rangle}$ if M satisfies irreducibility (i.e., the graph is strongly connected) and aperiodicity (i.e., the graph is non-bipartite). It is done by adding some probability to every element in M to ensure that it contains only positive probabilities. This makes every node connected to every other node, and thus guarantees that the graph is strongly connected and that it is not bipartite. A damping factor α and a reservoir of ranks represented by $U/|X_1|$ are introduced for this purpose. U is an $|X_1|$ by $|X_1|$ unit matrix and $U/|X_1|$ adds a weight to every edge uniformly. This gives us M' in Formula 4 and a new update rule as Formula 5.

$$M' = \alpha M + (1-\alpha) U/|X_1| \quad (4)$$

$$R_{X_1}^{\langle h+1\rangle} \leftarrow M' R_{X_1}^{\langle h\rangle} \quad (5)$$

Finally, our MedRank is the stationary ranking distribution $R_{X_1}^{\langle *\rangle}$ of type X_1 objects.

Algorithm 2 MedRank

Input: q: a queried disease; N: the extracted medical network; A: a list of attributes X_1, X_2, ..., X_n as ranking criteria, where the target attribute is X_1; K: the number of top-ranked elements; ϵ: a user defined threshold to determine convergence
Output: F: a ranked list of top-K treatments

1: $N' = \langle V', E'\rangle \leftarrow extractSubNetwork(N, q)$
2: Let $W_{X_t C}$ be matrices defined in the same way as in Formula 2
3: Initialize $R_{X_1} \leftarrow 1/|X_1|$.
4: **repeat**
5: $\quad newR_{X_1} \leftarrow (\alpha(\prod_{t=1}^{n-1} W_{X_t C} D_{CX_{t+1}}^{-1} W_{CX_{t+1}}) W_{X_n C} D_{CX_1}^{-1} W_{CX_1} + (1-\alpha) U/|X_1|) R_{X_1}$
6: $\quad difference \leftarrow |newR_{X_1} - R_{X_1}|$
7: $\quad R_{X_1} \leftarrow newR_{X_1}$
8: **until** $difference < \epsilon$ {until no change}
9: $F \leftarrow topK(R_{X_1}, K)$
10: **return** F

4.3 MedRank Algorithm

The proposed MedRank is presented as Algorithm 2. The algorithm uses a disease name q as a filter to extract a disease sub-network $N' = \langle V', E'\rangle$ from the medical information network N (line 1). q is currently limited to the terms included in the medical ontology, as the focus of the research is effective ranking. By default, X_1 is the target attribute of ranking interest. Firstly, the ranks of the target attribute objects are initialized uniformly based on the size of X_1 (line 3). With these initial ranks, the algorithm iteratively updates the ranks (REPEAT-UNTIL loop at lines 4-8) using Formula 5 defined earlier. The loop terminates at convergence, i.e., $|newR_{X_1} - R_{X_1}|$ is less than a threshold ϵ. Finally, the algorithm gets the top-K treatments (line 9) and returns F, a list of top-K ranked influential treatments.

The attribute types in set A are used as ranking criteria. They are used to find a set of objects belonging to these attribute types such that they are ranked high together in a disease sub-network. By computing eigenvector centrality among objects of these attribute types, it is to realize the idea that influential treatments are discussed by good articles, written by good author(s), published in a good journal, etc. As a treatment is always a treatment against a disease, it only makes sense to find influential treatment within the context of a disease. It will be shown later in Section 5.1.2 that the disease information can be found directly in MEDLINE records. Therefore, the key difference between MedRank and NetClus is that MedRank bases its ranking on the available category labels.

4.4 Computational Complexity Analysis

The time complexity of the network extraction phase is determined by a sequential scan on the records of MEDLINE for constructing a graph.

For the ranking on a disease sub-network $N' = \langle V', E'\rangle$, the time complexity for an iteration is $O(|E'|)$, where $|E'|$ is the number of edges in N'. This is because each link will be calculated at most twice in the matrix multiplication chain in Formula 2. For l iterations, the computational cost is $O(l|E'|)$.

5 Experiments and Evaluation

In this section we report the experiments and examine the effectiveness and efficiency of our MedRank algorithm. Five diseases are used, namely AIDS, Diabetes Mellitus Type II (D2), Hepatitis B (HB), Amyotrophic Lateral Sclerosis (ALS) and Rheumatoid Arthritis (RA). The selection of these diseases are based on the reasoning that for the commonly known diseases (i.e., AIDS, D2 and HB) our algorithm should be able to provide the obvious results as well as for the rarely known diseases (i.e., ALS and RA). In the evaluation, we compare MedRank with degree centrality and NetClus. Expert rankings for five diseases were collected from clinical professionals and aggregated, for each disease, into a consensus ranking for benchmarking. All experiments and evaluation are implemented in Visual Studio C# 2008 running on an Intel(R) Core(TM) i3 CPU laptop with a Windows 7 OS and a 4 GB RAM.

5.1 Data Sets

The data sets used in our experiments are the MEDLINE (2010) data set and MeSH (2010) ontology. It is to be noted that in this research, only the bibliographic information of medical literature is considered. The exploration of the rich information contained in article content is left as future work.

5.1.1 MeSH

Medical Subject Headings (MeSH) is a medical thesaurus and controlled semantic vocabulary that is part of the larger Unified Medical Language System (UMLS) thesaurus of NLM. It consists of a set of (57,229) *terms naming descriptors* that provide formal and explicit specifications of the present biomedical knowledge. Descriptors are arranged both alphabetically and hierarchically as a tree structure. Fig. 3 shows a top level view of the MeSH tree, where the disease category is expanded. Additional qualifiers, such as "Therapeutic Use", can be used to further categorize descriptors.

MeSH is chosen due to the convenience that MEDLINE records are indexed by it. The disease terms can be found in the "C" category. The treatment terms are in the following categories: "Therapeutics", "Anesthesia and Analgesia", "Surgical Procedures, Operative", and "Therapeutic Uses". All chemical substances that are labelled by the qualifier "Therapeutic Use" are also considered as treatments.

5.1.2 MEDLINE

MEDLINE is the premier bibliographic database of NLM. The data set is freely downloadable in XML format with 10GB in size (compressed) from NLM.[8] Each MEDLINE record is a reference to an article. As shown in Fig. 4, it contains the bibliographical information about the article, such as article ID (⟨*PMID*⟩), title (⟨*Article Title*⟩), author list (⟨*AuthorList*⟩) and journal title (⟨*Title*⟩). Further information, such as major diseases and treatments this article is about and which one of the four clinical trial phases (I to IV) the experiments are successful, is also available through the record's referenced MeSH ontology entries (⟨*MeshHeadingList*⟩). Topics being identified as relevant to the article are called descriptors. They are stored under the ⟨*DescriptorName*⟩ tag. The value of

[8] available at http://mbr.nlm.nih.gov/Download/index.shtml

1. + Anatomy [A]
2. + Organisms [B]
3. − Diseases [C]
   - Bacterial Infections and Mycoses [C01] +
   - Virus Diseases [C02] +
   - Parasitic Diseases [C03] +
   - Neoplasms [C04] +
   - Musculoskeletal Diseases [C05] +
   - ⋮
   - Wounds and Injuries [C26] +
4. + Chemicals and Drugs [D]
5. + Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]
6. + Psychiatry and Psychology [F]
7. + Phenomena and Processes [G]
8. + Disciplines and Occupations [H]
9. + Anthropology, Education, Sociology and Social Phenomena [I]
10. + Technology, Industry, Agriculture [J]
11. + Humanities [K]
12. + Information Science [L]
13. + Named Groups [M]
14. + Health Care [N]
15. + Publication Characteristics [V]
16. + Geographicals [Z]

Figure 3: Top level view of MeSH tree structure

```
<MedlineCitation Owner="NLM" Status="MEDLINE">
    <PMID Version="1">17713168</PMID>
    <Article PubModel="Print">
        <Journal>
            <ISSN IssnType="Print">1359-6535</ISSN>
            <JournalIssue CitedMedium="Print">
                <Volume>12</Volume>
                <Issue>5</Issue>
                <PubDate>
                    <Year>2007</Year>
                </PubDate>
            </JournalIssue>
            <Title>Antiviral therapy</Title>
                ⋮
        </Journal>
        <ArticleTitle>Declining prevalence of HIV-1 drug res
                ⋮
        <AuthorList CompleteYN="Y">
            <Author ValidYN="Y">
                <LastName>Di Giambenedetto</LastName>
                <ForeName>Simona</ForeName>
                <Initials>S</Initials>
            </Author>
                ⋮
    <MeshHeadingList>
        <MeshHeading>
            <DescriptorName MajorTopicYN="N">Anti-HIV Agents
            <QualifierName MajorTopicYN="Y">therapeutic use<
        </MeshHeading>
                ⋮
</MedlineCitation>
```

Figure 4: MEDLINE record example in XML

"MajorTopicYN" indicates whether this topic is considered as a major topic of the article.

5.2 Experiments

The ranking criteria are selected as {Author, Journal, Treatment, Clinical Trial}, because our basic assumption is that a *good* treatment is likely to be found in a *good* medical article published in a *good* journal, written by *good* author(s) and successful in clinical trials. Major diseases, treatments and clinical trials related to an article can be identified through MeSH descriptors (see Section 5.1.2). As in most cases, it is expected that a drug has to be successful in Phase III clinical trial in order to obtain approval from the appropriate regulatory agencies such as FDA (USA) or the EMA (European Union),[9] articles that are in a phase below III are excluded in our ranking. Thus, only the clinical trials under "Clinical Trial, Phase

[9] see http://en.wikipedia.org/wiki/Clinical_trial

Table 1: Size of sub-networks in categories

Type\N'	ALS	HB	AIDS	D2	RA
Article	7975	33679	48962	50732	70736
Author	16637	67320	86481	99060	108234
Journal	1256	2936	4272	3308	3963
Treatment	383	669	937	1121	1401
ClinicalTrial	5	5	5	5	5
Total	25256	104609	140657	154316	184339

Table 2: Top 10 influential treatments for AIDS

	Top 10 Treatments	Ranking
1	Zidovudine/therapeutic use	0.1500
2	Anti-HIV Agents/therapeutic use	0.1134
3	Antiretroviral Therapy, Highly Active	0.0855
4	Antiviral Agents/therapeutic use	0.0655
5	Anti-Retroviral Agents/ therapeutic use	0.0215
6	Interferon Type I/therapeutic use	0.0147
7	Didanosine/therapeutic use	0.0121
8	Ganciclovir/therapeutic use	0.0102
9	Antineoplastic Combined Chemotherapy Protocols/therapeutic use	0.0101
10	HIV Protease Inhibitors/ therapeutic use	0.0092

Table 3: Top 10 influential treatments for D2

	Top 10 Treatments	Ranking
1	Hypoglycemic Agents/therapeutic use	0.1859
2	Insulin/therapeutic use	0.0824
3	Metformin/therapeutic use	0.0379
4	Thiazolidinediones/therapeutic use	0.0364
5	Diabetic Diet	0.0340
6	Sulfonylurea Compounds/ therapeutic use	0.0271
7	Glyburide/therapeutic use	0.0181
8	Antihypertensive Agents/ therapeutic use	0.0176
9	Thiazoles/therapeutic use	0.0141
10	Self Care	0.0135

III", "Clinical Trial, Phase IV", "Controlled Clinical Trial", "Multicenter Study" and "Randomized Controlled Trial" categories are considered.

Table 1 shows the sizes of five disease sub-networks in terms of different object types. It can be seen that the Author and Article attribute types dominate the size of every sub-network. The table also shows that the size of a disease sub-network is significantly smaller than that of the entire network of MEDLINE's 20 millions of articles.

By running Algorithm 2 with α and ϵ set to 0.85 and 0.00001 respectively, the top-10 influential treatments for five diseases are reported in Tables 2 to 6. It should be noted that the numbers in the ranking column are the limiting probabilities, so the sum of ranks for all treatments in a disease sub-network equals to 1. It can be seen that except for ALS, the top one or two treatments of all other diseases are ranked significantly higher than the rest. Also, the differences between subsequent ranks in the lists decrease with the increase of their positions. These observations can be explained in two ways: Firstly, there are often one or two treatments more widely used against a disease though with exceptions (e.g., ALS). Secondly, this kind of eigenvector centrality-based method may not be sensitive enough to differentiate the order of less important items.

Table 4: Top 10 influential treatments for HB

	Top 10 Treatments	Ranking
1	Antiviral Agents/therapeutic use	0.1883
2	Lamivudine/therapeutic use	0.0915
3	Liver Transplantation	0.0602
4	Interferon-alpha/therapeutic use	0.0419
5	Interferon Type I/therapeutic use	0.0381
6	Reverse Transcriptase Inhibitors	0.0363
7	Interferons/therapeutic use	0.0295
8	Vaccination	0.0292
9	Interferon Alfa-2b/therapeutic use	0.0279
10	Phosphonic Acids/therapeutic use	0.0201

Table 5: Top 10 influential treatments for ALS

	Top 10 Treatments	Ranking
1	Neuroprotective Agents/ therapeutic use	0.0576
2	Riluzole/therapeutic use	0.0539
3	Antioxidants/therapeutic use	0.0326
4	Insulin-Like Growth Factor I/ therapeutic	0.0320
5	Respiration, Artificial	0.0295
6	Activities of Daily Living	0.0280
7	Thyrotropin-Releasing Hormone/ therapeutic use	0.0246
8	Excitatory Amino Acid Antagonists	0.0239
9	Creatine/therapeutic use	0.0239
10	Positive-Pressure Respiration	0.0218

Table 6: Top 10 influential treatments for RA

	Top 10 Treatments	Ranking
1	Antirheumatic Agents/ therapeutic use	0.2420
2	Antibodies, Monoclonal/ therapeutic use	0.0709
3	Methotrexate/therapeutic use	0.0546
4	Anti-Inflammatory Agents/therapeutic	0.0303
5	Anti-Inflammatory Agents, Non-Steroidal	0.0266
6	Sulfasalazine/therapeutic use	0.0160
7	Penicillamine/therapeutic use	0.0156
8	Gold Sodium Thiomalate/ therapeutic use	0.0131
9	Glucocorticoids/therapeutic use	0.0116
10	Immunosuppressive Agents/therapeutic	0.0103

5.3 Evaluation

This section reports the evaluation of the proposed MedRank algorithm. It includes the aggregation of expert consensuses from collected expert rankings and the comparisons of MedRank against two baseline methods.

5.3.1 Expert Consensus Aggregation

Since there is no *ground truth* for evaluation, expert opinions are used as an alternative. They are aggregated into a consensus ranking for each of the five diseases for benchmarking.

Expert Ranking Collection We distributed 1,500 questionnaires to the experts of state hospitals and medical research institutions over four countries, and received 106 valid responses from five hospitals in mainland China and Taiwan. System ranked top-10

Table 7: Expert feedback size for five diseases

AIDS	D2	HB	ALS	RA
24	28	21	16	17

most influential treatments for five diseases are listed respectively in the questionnaire. Participants were asked to answer on only the diseases they were familiar with. They were asked to provide their own ranking lists for a disease, if they did not agree with the system ranking. The sizes of the collected feedback are shown in Table 7.

Concordance Measurements For the purpose of evaluation, we need to measure if two ranking lists, say S and T, are concordant or agree with each other. This is the problem of comparing *partial rankings* (Fagin et al. 2004), i.e., one list may contain elements not found in the other list, as questionnaire participants had the option to provide their own lists with items not in the system ranking. *Total rankings* are studied in classical rank correlation methods such as Kendall's τ (Kendall 1948) and Spearman's ρ (Spearman 1904). Webber et al. (2010) have classified measurements that compare two lists into four categories based on whether a measurement is applicable to partial rankings and whether it considers top-weightedness, i.e., the top of the list is weighted higher than the tail.

Fagin et al. (2004) proposed several measurements for top-k partial rankings. We choose their extension of Kendall's τ to represent the class of measurements that are not top-weighted. The method compares every pair of elements in the union set of elements appear in the ranking lists. There are four cases to consider for penalizing displacements. The sum of all penalties normalized by the number of all possible pairwise comparisons gives a degree p of displacement in the [0,1] interval. We define Fagin's τ as $1 - p$ for measuring concordance. However, the method disregards the position where the displacement occurs.

As Webber et al. (2010) argues that the top of the list is often considered more important than the tail in top-K rankings, we adopt the intersection metric of Fagin et al. (2004) (referred as AO) that captures this top-weightedness as the main measurement.

$$AO(S,T,K) = \frac{1}{K}\sum_{d=1}^{K}\frac{|S_{:d} \cap T_{:d}|}{d} \qquad (6)$$

where K is the top-K ranked items of interest; S and T are two ranking lists; $S_{:d}$ denotes the set of elements in S from the first up to the d^{th} position; $T_{:d}$ is defined in the same way.

AO is the average over the sum of the weighted overlaps of the first d elements in both lists. The score is in the interval of [0,1], where 0 means no items shared by two lists and 1 means two lists are identical. It can be seen from Formula 6 that the weight up to the d^{th} element decreases with the increase of d. Thus, it is top-weighted. Also, AO is lenient to a displacement occurring at two very close positions, say the 2^{nd} position in list S and 3^{rd} position in list T, but harsh to those that are far apart (as the score will not be granted until the item has been found in both lists at the d^{th} position).

Expert Agreement It is important to measure the compactness or degree of agreement among expert rankings, because it makes more sense to find a consensus among them if they tend to agree with each other. The compactness is calculated using Formula 7 adapted from (Xiaoyun et al. 2009). In the formula, x_i and x_j are any two lists and m is the number of all ranking lists. It calculates the average over the pairwise distance (using AO) between ranking lists. As AO is a measurement of concordance, the higher the pairwise AO is, the higher the overall concordance would be.

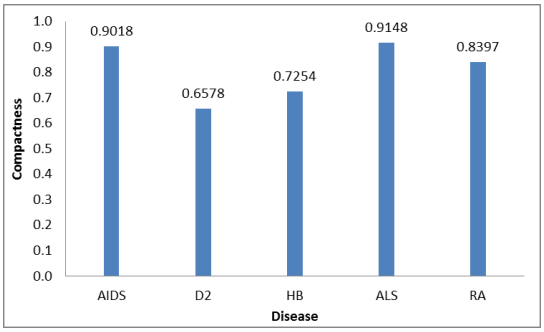


Figure 6: Agreements among expert rankings for five diseases

Table 8: Aggregated Expert Ranking

Disease	Aggregated List
AIDS	1,2,3,4,5,6,7,8,10,9
D2	1,5,3,6,2,4,10,7,8,9
HB	1,2,4,5,6,3,7
ALS	1,2,4,3,8,5,6,9,10,7
RA	1,2,3,4,5,6,7,9,8,10

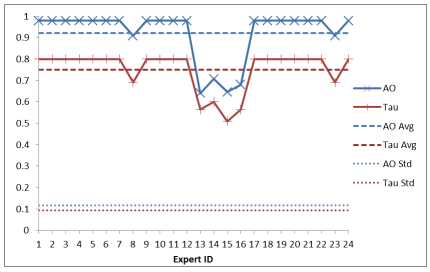
$$compactness = \sqrt{\frac{\sum_{i=1}^{m}\sum_{j=1}^{m}(x_i - x_j)^2}{m(m-1)}} \qquad (7)$$

The results are illustrated in Fig 6. The figure shows that the degrees of agreement among expert rankings are relatively low in Diabetes Mellitus Type II (66%) as well as Hepatitis B (73%). It can also be seen that the degrees for the other three diseases are relatively high (above 84%). These results indicate that the aggregated expert consensuses for AIDS, ALS and RA are more indicative than D2 and HB for evaluating MedRank and baseline methods.

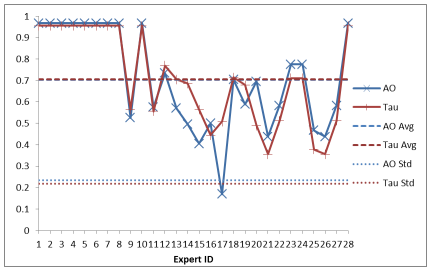
Ranking Aggregation Expert ranking lists are aggregated into a consensus list per disease. We adopt a *2-approximation* aggregation algorithm proposed by Chin et al. (2004). This heuristic algorithm constructs a single ranking from a list of *partial rankings* with respect to the maximization of the consensus. Table 8 shows the aggregated expert rankings for five diseases, where integer denotes the position of a treatment originally in the system ranking. It is to be noted that for Hepatitis B, three treatments have been removed from all the rankings (of the system and baselines) in the evaluation. This is caused by the sub/super MeSH categories that are considered as the same treatment approach by the experts.

5.3.2 MedRank vs. Expert Rankings

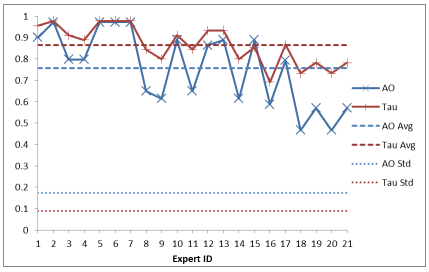
System rankings have been first evaluated, for each disease, against expert rankings one by one to give a sense of their concordance with individual expert opinions. The results are illustrated in Fig. 5 (a)-(e). The average and standard deviation of AO and Fagin's τ are presented as doted lines. Generally, system rankings have reasonably high concordance with



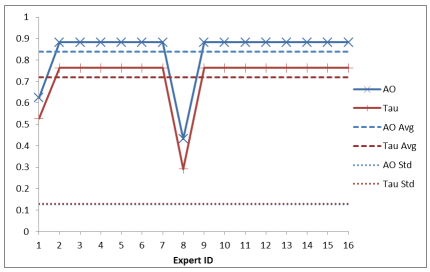
(a) AIDS



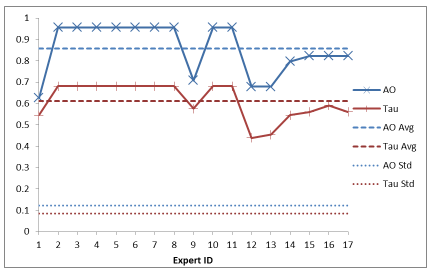
(b) D2



(c) HB



(d) ALS



(e) RA

Figure 5: MedRank vs. every expert ranking for five diseases

expert rankings, with the average *AO* and Fagin's τ above 0.7 except for Fagin's τ in RA. The results are especially good for AIDS, Amyotrophic Lateral Sclerosis (ALS) and Rheumatoid Arthritis (RA), where average *AO*s are above 0.83 with standard deviation around 0.1. Also, for these three diseases, *AO* scores are almost all higher than Fagin's τ. This suggests that most displacements occur close to the tail of the list or just a few positions away from where an item is supposed to be. By contrast, *AO* scores are mostly lower than Fagin's τ for Hepatitis B (HB). This suggests the opposite case. By looking at the aggregated expert ranking for HB in Table 8 presented earlier, we can see that the positions of Treatment item 3 are far apart in the system ranking and expert aggregated ranking (i.e., 3 vs. 6). As for Diabetes Mellitus Type II (D2), it is a mixture of above two cases. Overall, MedRank gives pretty good rankings for AIDS, ALS and RA.

5.3.3 MedRank vs. Baselines

We report the evaluation of our MedRank algorithm against two baseline methods, i.e., degree centrality and NetClus.

Baseline Settings Degree centrality counts, for every treatment in a disease sub-network, the degree of the treatment node. It is equivalent to frequency counting, i.e., counting the number of articles linked to the treatment. From these frequencies, a ranking list of treatments can be obtained. The concordance between the ranking lists and expert consensuses is measured by *AO* and Fagin's τ. They are presented as *Degree Centrality* in Fig. 7 and Fig. 8 respectively.

As NetClus is a ranking-based clustering algorithm, we have experimented on the cases of $k = 2$ and $k = 5$, where k is the number of clusters. A 5-disease sub-network of the same five diseases used for MedRank is extracted for the case of $k = 5$ and all possible 2-disease sub-networks selected from these five diseases are extracted for the case of $k = 2$. In addition to our ranking criteria, other attribute types, such as "Disease" and "Term" (extracted from article title with stemming (Porter 1980)) have been tried to help clustering. Best results have been obtained by

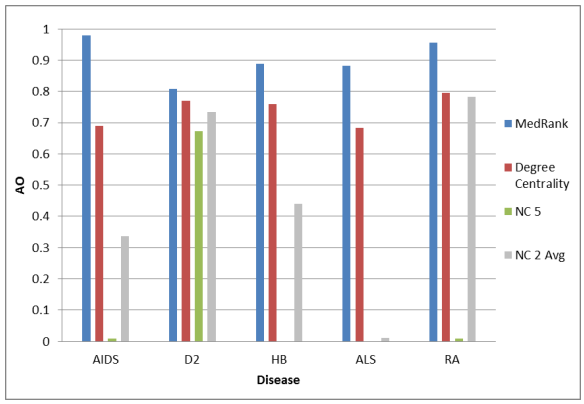


Figure 7: MedRank vs. baselines using *AO*

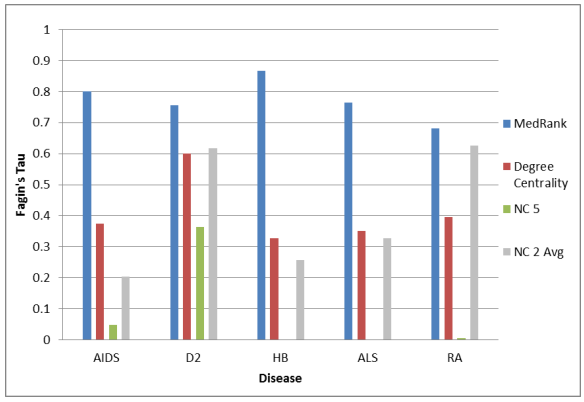


Figure 8: MedRank vs. baselines using Fagin's Tau

using "Term" with prior probabilities (as used also by Sun, Yu & Han (2009)). Terms have been chosen based on their degree centrality with common terms such as "patient" and "disease" removed. The results for the 5-disease sub-network is presented as *NC 5* in Fig. 7 and Fig. 8, while *NC 2 Avg* denotes the average concordance of a disease over all 2-disease sub-networks that contain the disease.

Analysis From Fig. 8 we can see that degree centrality has relatively low Fagin's τ scores (mostly about half of MedRank's scores) but better *AO* scores

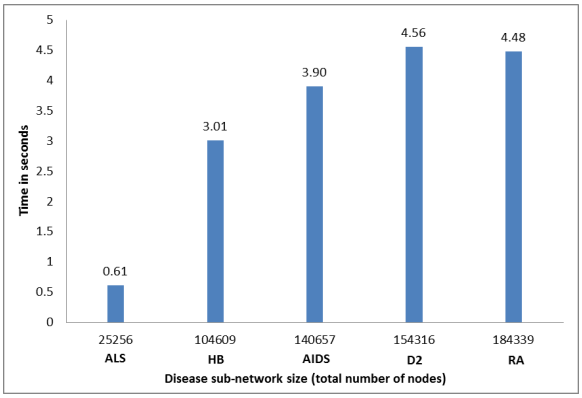


Figure 9: Time spent on the ranking for five sub-networks

in Fig. 7. This suggests that its outputs have many displacements but the portion occurring at the top of the lists is not high. When $k = 5$ NetClus is not able to give good clusters, as D2 treatments dominate the rankings. Thus only the ranking in the D2 cluster has high concordance score in Fig. 7 and Fig. 8. NetClus performs slightly better when $k = 2$, though it can still be seen from both figures that D2 and RA treatments dominate the ranking. Overall, the figures show that MedRank outperforms degree centrality and NetClus.

6 Discussions

In this section we discuss the effectiveness, efficiency and implications of MedRank.

6.1 Effectiveness and Efficiency

Section 5.3 presented how our system rankings are evaluated against the "consensus" rankings aggregated from the expert rankings and the baseline methods. Figs. 7 and 8 have shown that the proposed network-based ranking algorithm MedRank is effective as it outperforms the baselines.

We include a chart in Fig. 9 to show the scalability and efficiency of the ranking phase. The sub-networks corresponding to the selected five diseases are ordered according to their sizes, i.e., total number of nodes (Table 1). The times spent on ranking the sub-networks using MedRank are measured in seconds. It can be seen from this empirical analysis that the computational cost for the ranking phase is linear.

6.2 Implications of MedRank

As a few doctors mentioned in their feedback that although there may be consensus among clinical professionals on what treatment to be applied to a disease (e.g., the "cocktail" method for AIDS), whether a treatment is better than another should be judged case by case. This goes in line with the motivation of our research. Our intention is not to compete with the best medical experts in giving clinical advices but to show a methodology like this can provide ranking on the influential treatments. Even if the ranking may not always be the most authoritative one, it provides much value in showing information technology can efficiently filter out noises and derive highly valued candidate treatments for further study.

This study has been focused on the medical domain based on tagged medical literature. MedRank presents an interesting methodology for ranking an information network based on the available category labels (which is different from RankClus and NetClus), as well as their associated, multi-type semi-structured entities (which is different from PageRank and PopRank). Therefore, it represents a new and interesting method for ranking categorized entities in multi-dimensional information networks. The success of categorized entity ranking in medical literature domain suggests a new methodology and a potential success in other domains, as long as entities and their relationships can be identified. This opens an interesting direction for further study.

7 Conclusions

In a general medical information network, objects such as patients, clinical trials, symptoms, diseases, medical journal articles, or treatments, can all be linked together through different kinds of referential relationships. Then by applying a network-based ranking algorithm, we can use a query mechanism such as the *slash-tag* search engine[10] to search for top-ranked objects according to their categories (i.e., tags). In this paper we demonstrated a pioneer research for ranking treatments for given diseases based on a medical information network.

The contribution of this research is threefold. Firstly, we extracted heterogeneous objects from medical literature as an information network for medical knowledge management. Secondly, we proposed a new network-based ranking algorithm, namely MedRank, to rank the most influential treatments. Thirdly, we successfully conducted a survey with clinical practitioners to collect expert rankings for benchmarking. The proposed algorithm has been evaluated against two baseline methods. It has been shown that MedRank is effective and efficient.

For future research, we will extend this network-based ranking approach to other domains. Investigations on ranking emerging medical treatments for new and unknown diseases from medical literature will also be considered.

8 Acknowledgements

We express our gratitude to Professor Len Gray (School of Medicine, University of Queensland) for his insightful comments and discussions on this research work. We thank Dr. Chiu, Allen Wen-Hsiang, Dr. Liu, Chi-chun, and Mr. Liang, Jun for organising questionnaire distribution and data collection. We also thank the doctors from the First Affiliated Hospital and the Second Affiliated Hospital of Medical School of Zhejiang University, the First People's Hospital of Hangzhou, the Taipei Medical University Hospital, and the Taipei City Hospital for their valuable feedback.

This work was supported in part by the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053(NS-CTA), US NSF IIS-0905 215 and IIS-1017362, and U.S. AFSOR MURI award FA9550-08-1-0265.

References

Adomavicius, G. & Tuzhilin, A. (2005), 'Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions', *IEEE Transactions on Knowledge and Data Engineering* **17**(6), 734 – 749.

[10] http://blekko.com/

Berner, E. S. (2007), *Clinical decision support systems: theory and practice*, Springer, New York.

Borodin, A., Roberts, G., Rosenthal, J. & Tsaparas, P. (2005), 'Link analysis ranking: algorithms, theory, and experiments', *ACM Transactions on Internet Technology (TOIT)* **5**(1), 231–297.

Ceci, M., Appice, A., Loglisci, C. & Malerba, D. (2010), Complex objects ranking: a relational data mining approach, *in* 'Proceedings of the 2010 ACM Symposium on Applied Computing', ACM, pp. 1071–1077.

Chen, R., Lin, H. & Yang, Z. (2011), 'Passage retrieval based hidden knowledge discovery from biomedical literature', *Expert Syst. Appl.* **38**, 9958–9964.

Chin, F., Deng, X., Fang, Q. & Zhu, S. (2004), 'Approximate and dynamic rank aggregation', *Theoretical computer science* **325**(3), 409–424.

Easley, D. & Kleinberg, J. (2010), *Networks, crowds, and markets: Reasoning about a highly connected world*, Cambridge Univ Pr.

Fagin, R., Kumar, R. & Sivakumar, D. (2004), 'Comparing top k lists', *SIAM Journal on Discrete Mathematics* **17**(1), 134–160.

Hliaoutakis, A., Zervanou, K. & Petrakis, E. (2009), 'The amtex approach in the medical document indexing and retrieval application', *Data & Knowledge Engineering* **68**(3), 380–392.

Kendall, M. (1948), *Rank Correlation Methods*, Charles Griffin & Co.

Kleinberg, J. (1999), 'Authoritative sources in a hyperlinked environment', *Journal of the ACM (JACM)* **46**(5), 604–632.

Langville, A. & Meyer, C. (2004), 'Deeper inside pagerank', *Internet Mathematics* **1**(3), 335–380.

Luo, G. (2009), Design and evaluation of the iMed intelligent medical search engine, *in* 'IEEE 25th international conference on Data engineering (ICDE'09)', pp. 1379 –1390.

Luo, G. & Tang, C. (2008), On iterative intelligent medical search, *in* 'Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'08)', ACM, pp. 3–10.

Luo, G., Tang, C., Yang, H. & Wei, X. (2008), Medsearch: a specialized search engine for medical information retrieval, *in* 'Proceeding of the 17th ACM conference on Information and knowledge management (CIKM'08)', ACM, pp. 143–152.

Munteanu, C. R., Magalhâes, A. L., Uriarte, E. & González-Díaz, H. (2009), 'Multi-target qpdr classification model for human breast and colon cancer-related proteins using star graph topological indices', *Journal of Theoretical Biology* **257**(2), 303 – 311.

Newman, M. (2010), *Networks: an introduction*, Oxford Univ Pr.

Nie, Z., Zhang, Y., Wen, J. & Ma, W. (2005), Object-level ranking: Bringing order to web objects, *in* 'Proceedings of the 14th international conference on World Wide Web (WWW'05)', ACM, pp. 567–574.

Page, L., Brin, S., Motwani, R. & Winograd, T. (1999), The pagerank citation ranking: Bringing order to the web., Technical Report 1999-66, Stanford InfoLab.

Petric, I., Urbancic, T., Cestnik, B. & Macedoni-Luksic, M. (2009), 'Literature mining method rajolink for uncovering relations between biomedical concepts', *Journal of Biomedical Informatics* **42**(2), 219–227.

Porter, M. (1980), 'An algorithm for suffix stripping', *Program* **14**(3), 130–137.

Ratprasartporn, N., Po, J., Cakmak, A., Bani-Ahmad, S. & Ozsoyoglu, G. (2009), 'Context-based literature digital collection search', *The VLDB Journal* **18**, 277–301.

Shardanand, U. & Maes, P. (1995), Social information filtering: algorithms for automating "word of mouth", *in* 'Proceedings of the SIGCHI conference on Human factors in computing systems (CHI'95)', ACM, pp. 210–217.

Spearman, C. (1904), 'The proof and measurement of association between two things', *The American journal of psychology* **15**, 72–101.

Sun, Y., Han, J., Zhao, P., Yin, Z., Cheng, H. & Wu, T. (2009), Rankclus: integrating clustering with ranking for heterogeneous information network analysis, *in* 'Proceedings of the 12th international conference on Extending database technology (EDBT'09)', pp. 565–576.

Sun, Y., Yu, Y. & Han, J. (2009), Ranking-based clustering of heterogeneous information networks with star network schema, *in* 'Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'09)', ACM, pp. 797–806.

Webber, W., Moffat, A. & Zobel, J. (2010), 'A similarity measure for indefinite rankings', *ACM Transactions on Information Systems (TOIS)* **28**(4), 20.

Xiaoyun, C., Yi, C., Xiaoli, Q., Min, Y. & Yanshan, H. (2009), Pgmclu: A novel parallel grid-based clustering algorithm for multi-density datasets, *in* '1st IEEE Symposium on Web Society (SWS'09)', pp. 166 –171.

Yetisgen-Yildiz, M. & Pratt, W. (2006), 'Using statistical and knowledge-based approaches for literature-based discovery', *Journal of Biomedical Informatics* **39**(6), 600–611.

Zhao, D. & Weng, C. (2011), 'Combining pubmed knowledge and ehr data to develop a weighted bayesian network for pancreatic cancer prediction', *Journal of Biomedical Informatics* **44**(5), 859 – 868.