Title: Ranking Methods for Networks

Name: Yizhou Sun[1], Jiawei Han[2]

Affil./Addr. 1: College of Computer and Information Science, Northeastern University, Boston, MA, USA,
yzsun@ccs.neu.edu

Affil./Addr. 2: Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, USA,
hanj@cs.uiuc.edu

Ranking Methods for Networks

Synonyms

Importance Ranking; Identify Influential Nodes; Relevance Ranking; Link-based Ranking

Glossary

Ranking: sort objects according to some order.

Global ranking: objects are assigned ranks globally.

Query-dependent ranking: objects are assigned with different ranks according to different queries.

Proximity ranking: objects are ranked according to proximity or similarity to other objects.

Homogeneous information network: networks that contain one type of objects and one type of relationships.

Heterogeneous information network: networks that contain more than one type of objects and/or one type of relationships.

Learning to rank: ranking are learned according to examples via supervised or semi-supervised methods.

Definition

Ranking objects in a network may refer to sorting the objects according to importance, popularity, influence, authority, relevance, similarity, and proximity, by utilizing link information in the network.

Introduction

In this article, we introduce the ranking methods developed for networks. Different from other ranking methods defined in text or database systems, links or the structure information of the network are significantly explored. For most of the ranking methods in networks, ranking scores are defined in a way that can be propagated in the network. Therefore, the rank score of an object is determined by other objects in the network, usually with stronger influence from closer objects and weaker influence from more remote ones.

Methods for ranking in networks can be categorized according to several aspects, such as *global ranking* vs. *query-dependent ranking*, based on whether the ranking result is dependent on a query; *ranking in homogeneous information networks* vs. *ranking in heterogeneous information networks*, based on the type of the underlying networks; *importance-based ranking* vs. *proximity-based ranking*, based on whether the semantic meaning of the ranking is importance related or similarity/promximity related; and *unsupervised* vs. *supervised* or *semi-supervised*, based on whether training is needed.

Historical Background

The earliest ranking problem for objects in a network was proposed by sociologists, who introduced various kinds of centrality to define the importance of a node (or actor) in a social network. With the advent of the World-Wide Web and the rising necessity of Web search, ranking methods for Web page networks are flourishing, including the well known ranking methods, PageRank [Brin and Page(1998)] and HITS [Kleinberg(1999)]. Later, in order to better support entity search instead of Web page ranking, object ranking algorithms are proposed, which usually consider more complex structural information of the network, such as heterogeneous information networks. Moreover, in order to better personalize search quality, ranking methods that can integrate user guidance are proposed. Learning to rank techniques are used in such tasks, and not only the link information but the attributes associated with nodes and edges are commonly used.

Methods and Algorithms

In this section, we introduce the most representative ranking methods for networks.

Centrality and Prestige

In network science, various definitions and measures are proposed to evaluate the prominence or importance of a node in the network. According to [Wasserman and Faust(1994)], centrality and prestige are two concepts to quantify prominence of a node within a network, where *centrality* focuses on evaluating the involvement of a node no matter whether the prominence is due to the receiving or the transmission of the ties, whereas *prestige* focuses on evaluating a node according to the ties that the node is receiving.

Given a network $G = (V, E)$, where V and E denote the vertex set and the edge set, several frequently used centrality measures are listed in the following.

- **Degree Centrality.** Degree centrality [Nieminen(1974)] of a node u is defined as the *degree* of nodes in the network: $C_D(u) = \sum_v A_{u,v}$, where A is the adjacency matrix of G. *Normalized degree* $C'_D(u) = C_D(u)/(N-1)$ can also be used to measure the relative importance of a node, where N is the total number of nodes in the network, and $N-1$ is the maximum degree that a node can have.
- **Closeness Centrality.** *Closeness centrality* [Sabidussi(1966)] assigns a high score to a node if it is close to many other nodes in the network, and is calculated by the inverse of the sum of geodesic distance (shortest distance) between the node and other nodes:

$$C_C(u) = \frac{1}{\sum_v d(u,v)}$$

where $d(u, v)$ is the geodesic distance between u and v. A *normalized closeness centrality score* [Beauchamp(1965)] is defined as:

$$C'_C(u) = \frac{N-1}{\sum_v d(u,v)}$$

where $N-1$ is the possible minimum sum of distances between a node and the remaining $N-1$ nodes.

- **Betweenness Centrality.** *Betweeness centrality* evaluates how many times the node falls on the shortest or geodesic paths between a pair of nodes:

$$C_B(u) = \sum_{v<w} \frac{g_{vw}(u)}{g_{vw}}$$

where g_{vw} is the number of shortest paths between v and w, and $g_{vw}(u)$ is the number of shortest paths between v and w containing u. A *normalized betweenness centrality score* is given in [Freeman(1977)]:

$$C'_B(u) = \frac{2C_B(u)}{N^2 - 3N + 2}$$

where $(N^2-3N+2)/2$ can be proved to be the maximum value of $C_B(u)$, when u is a center point in a star network.

The readers may refer to [Freeman(1978)] and [Wasserman and Faust(1994)] for detailed introduction of these centrality measures.

In [Wasserman and Faust(1994)], several prestige measures are proposed for directed networks.

- **Degree Prestige.** *Degree prestige* is defined as the in-degree of each node, as a node is prestigious if it receives many nominations:

$$P_D(u) = d_{in}(u) = \sum_v A_{v,u}$$

The *normalized version of degree prestige* is:

$$P'_D(u) = \frac{P_D(u)}{N-1}$$

where N is the total number of nodes in the network, and thus $N-1$ is the maximum in-degree that a node can have .

- **Eigenvector-based Prestige.** In order to capture the intuition that a node is prestigious if it is linked by a lot of prestigious nodes. *Eigenvector-based prestige* is proposed in an iterative form:

$$P(u) = \frac{1}{\lambda} \sum_v A_{v,u} P(v)$$

It turns out that $\mathbf{p} = (P(1), \ldots, P(N))'$ is the primary eigenvector of the transpose of adjacency matrix A^T. $\mathbf{p}$ is also called *eigenvector centrality*.

- **Katz Prestige.** In [Katz(1953)], attenuation factor α is considered for influence with longer length transmissions, and the *Katz score* is calculated as a weighted combination of influence with different lengths:

$$P_{Katz}(u) = \sum_{k=1} \alpha^k \sum_v (A^k)_{vu}$$

which can be written into the matrix from:

$$\mathbf{P_{Katz}} = ((I - \alpha A)^{-1} - I)'\mathbf{1}$$

where $\mathbf{P_{Katz}} = (P_{Katz}(1), \ldots, P_{Katz}(N))'$, I is the identity matrix, and $\mathbf{1}$ is an all-one vector with length N. Katz score is also called Katz centrality.

Global Ranking

Along with the flourish of Web applications, many link-based ranking algorithms are proposed. We first introduce the ranking algorithms that assign global ranking scores to objects in the network.

PageRank

In information network analysis, the most well-known ranking algorithm is PageRank [Brin and Page(1998)], which has been successfully applied to the Web search problem. PageRank is a link analysis algorithm that assigns a numerical weight to each object in the information network, with the purpose of "measuring" its relative importance within the object set.

More specifically, for a directed web page network G with adjacency matrix A, the PageRank rank score of a web page u is iteratively determined by the scores of its incoming neighbors:

$$PR(u) = \frac{1-\alpha}{N} + \alpha \sum_{v} A_{vu} PR(v)/d_{out}(v)$$

where $\alpha \in (0, 1)$ is a damping factor and is set as 0.85 in the original PageRank paper, N is the total number of nodes in the network, and $d_{out}(v) = \sum_w A_{vw}$ is the degree of out-going links of v. The iterative formula can also be written in the following matrix form:

$$\mathbf{PR} = \frac{1-\alpha}{N}\mathbf{1} + \alpha M^T \mathbf{PR}$$

where M is the row normalized matrix of A, i.e., $M_{uv} = A_{uv}/\sum_{v'} A_{uv'}$, and $\mathbf{1}$ is an all-one vector with length N.

The iterative formula can be proved to converge to the following stable point:

$$\mathbf{PR} = (I - \alpha M^T)^{-1}\frac{1-\alpha}{N}\mathbf{1},$$

where I is the identity matrix.

PageRank score can be viewed as a stationary distribution of a random walk on the network, where a random surfer either randomly selects an out-linked web page v of the current page u with probability $\alpha/d_{out}(u)$, or randomly selects a web page from the whole web page set with probability $(1-\alpha)/N$.

Query-Dependent Ranking

Different from global ranking, query-dependent ranking produces different ranking results for different queries.

HITS

Hyperlink-Induced Topic Search (HITS) [Kleinberg(1999)] ranks objects based on two scores: *authority* and *hub*. Authority estimates the value of the content of the object, whereas hub measures the value of its links to other objects.

HITS is designed to be applied on a query dependent subnetwork, where the most relevant (e.g., by keyword matching) web pages to the query are first extracted. Then the authority and hub scores are calculated according to the following two rules:

1. An object has a high authority score if it is pointed by many nodes with high hub scores; and
2. An object has a high hub score if it has pointed to many nodes with high authority scores.

Mathematically, the two rules can be represented as two formulas:

$$Auth(u) = \sum_v A_{vu} Hub(v)$$

$$Hub(u) = \sum_v A_{uv} Auth(v)$$

where A is the adjacency matrix of the subnetwork. The two formulas are calculated iteratively, where normalization is needed after each iteration such that the score summation for each type equals to 1.

By reforming the two formulas into matrix form, we can find the authority score vector is the primary eigenvector of $A^T A$ matrix, and the hub score vector is the primary eigenvector of AA^T matrix.

Note that the authority and hub scores can only be calculated at query time, as the subnetwork needs first be extracted according to the query. Therefore, efficiency is a major issue of the HITS algorithm.

Topic-Sensitive PageRank

In order to obtain both the offline computation benefit as PageRank and the query-dependent ranking benefit as HITS, topic-sensitive PageRank is proposed in [Haveliwala(2002)].

The topic-sensitive PageRank is comprised of two steps. In Step 1, a biased PageRank score vector is computed for each predefined topic offline; and in Step 2, the probabilities that a query belongs to each topic are determined online, and the final query-dependent ranking is a weighted combination of the rankings for each topic.

More specifically, in Step 1, let T_j be the web page set for topic c_j, and let $\mathbf{p}_j$ be the initial ranking score vector for topic c_j, where $p_j(u) = 1/|T_j|$ if web page $u \in T_j$ and $P_j(u) = 0$ otherwise, the biased PageRank score for topic c_j is calculated as:

$$\mathbf{PR}_j = (1-\alpha) M^T \times \mathbf{PR}_j + \alpha \mathbf{p}_j$$

where M is the row normalized matrix of adjacency matrix A, as defined in PageRank section, and α is the parameter indicating the weight for the initial ranking vector.

Note that, in PageRank, the initial ranking score is $1/N$ for all the web pages in the network.

In Step 2, for a given query q, the probability it belongs to each topic c_j is calculated according to the term distribution in each topic:

$$P(c_j|q) \propto P(c_j)P(q|c_j)$$

where $P(c_j)$ is the prior distribution of topic c_j, and $P(q|c_j)$ is the probability that query q can be generated in topic c_j according to term distribution in c_j. Then, the query q dependent importance score for web page u can be calculated as:

$$s_{qu} = \sum_j P(c_j|q)PR_j(u)$$

where $PR_j(u)$ is the biased PageRank score for web page u for topic c_j.

Personalized PageRank

In [Jeh and Widom(2003)], personalized PageRank is proposed and how to scale the computation is introduced. Personalized PageRank aims at calculating biased PageRank score to a personalized query vector $\mathbf{q}$, which is called preference vector:

$$\mathbf{PPR}_q = (1-\alpha)M^T \times \mathbf{PPR}_q + \alpha\mathbf{q}$$

where M is the row normalized matrix for the network, and $\alpha \in (0,1)$ is the parameter indicating the probability a random walk will teleport to the query vector. $\mathbf{PPR}_q$ is called the personalized PageRank vector (PPV) for preference vector $\mathbf{q}$.

Different from topic-sensitive PageRank, where the query vectors are fixed for predefined topics, query vectors in personalized PageRank are arbitrary. Therefore, how to compute personalized PageRank efficiently online becomes critical, and the readers may refer to [Jeh and Widom(2003)] for more discussions.

A similar idea, TrustRank, that is used for ranking web pages according to their trustability is proposed in [Gyöngyi et al(2004)Gyöngyi, Garcia-Molina, and Pedersen], where the query vector is determined by a set of carefully selected trustable web sites.

Ranking in Heterogeneous Information Networks

Traditional ranking problem is considered in homogeneous information networks, where the networks contain only one type of objects and the objects are connected via one type of relationships. Recently, ranking algorithms for heterogeneous information networks are proposed, where the networks contain multiple types of objects and/or multiple types of relationships.

ObjectRank

ObjectRank is proposed in [Balmin et al(2004)Balmin, Hristidis, and Papakonstantinou], which aims at ranking the objects according to a keyword-based query in a database. A database is represented using a *labeled data graph*, $D(V_D, E_D)$, where nodes represented objects from different types and links represented relationships from different types. A *schema graph*, $G(V_G, E_G)$, is used to describe the structure of the data graph. Each node also contains several attribute-value pairs, which determine a set of keywords each node is associated with.

An *authority transfer schema graph*, $G^A(V_G, E_G^A)$, is then defined according to the schema graph, where authority transfer rates are given to the edges in the schema graph, that is, a certain link type in the data graph. The rate is specified by domain experts or obtained by trial-and-error. Afterwards, an *authority transfer data graph*, $D^A(V_D, E_D^A)$, can be derived, where the authority transfer rate between two objects u and v is defined by:

$$M(u,v) = \begin{cases} \frac{w(T)}{d_{out}(u,T)} & \text{if } d_{out}(u,T) > 0 \\ 0 & \text{if } d_{out}(u,T) = 0 \end{cases}$$

where T is the type of edge $e = (u, v)$, $w(T)$ is the authority transfer rate on the type of edges T, and $d_{out}(u, T)$ is the total number of out edges from u and of type T. After defining the authority transfer data graph and obtaining the new transition matrix M defined on objects, the online query processing is similar to personalized PageRank. For a keyword query k, the system will prepare the query vector $\mathbf{q}$ according to the set of objects containing the keyword. If an object u contains the keyword, then $q(u) = 1/N_k$, where N_k is the total number of objects containing the keyword k; otherwise, $q(u) = 0$. Then the *ObjectRank vector* for objects given the keyword k is defined as:

$$\mathbf{OR}_q = (1-\alpha)M^T \times \mathbf{OR}_q + \alpha\mathbf{q}$$

where α is the parameter indicating the probability a random walk will teleport to the query vector.

PopRank

In [Nie et al(2005)Nie, Zhang, Wen, and Ma], PopRank is proposed to rank web objects by using both web links and object relationship links. The PopRank score vector $\mathbf{R}_X$ for objects from type X is defined as a combination of their Web popularity $\mathbf{R}_{EX}$ and impacts from objects from other types:

$$\mathbf{R}_X = \epsilon\mathbf{R}_{EX} + (1-\epsilon)\sum_Y \gamma_{YX} M_{YX}^T \mathbf{R}_Y$$

where ϵ is the weighting parameter of the two components, γ_{YX} is the *popularity propagation factor* (PPF) of the relationship link from an object of type Y to an object of type X and $\sum_Y \gamma_{YX} = 1$, M_{YX} is the row normalized adjacency matrix between type Y and type X, and $\mathbf{R}_Y$ is the PopRank score vector for type Y.

In the paper, a simulated annealing-based algorithm for learning popularity propagation factor γ_{YX} is also proposed, according to some partial ranking lists given by users. Note that, PopRank assigns a global score for every object.

Authority Ranking for Heterogeneous Bibliographic Network

In reality, ranking function is not only related to the link property of an information network, but also dependent on the hidden ranking rules used by people in some specific domain. Ranking functions should be combined with link information and user rules in that domain. *Authority ranking* for heterogeneous bibliographic network is proposed in [Sun et al(2009a)Sun, Han, Zhao, Yin, Cheng, and Wu] and [Sun et al(2009b)Sun, Yu, and Han], which gives an object higher rank score if it has more authority.

Without using citation information, as citation information could be unavailable or incomplete (such as in the DBLP data, where there is no citation information imported from Citeseer, ACM Digital Library, or Google Scholars), two simple empirical rules similar to HITS are proposed to rank authors and venues:

- Rule 1: Highly ranked authors publish *many* papers in highly ranked venues.
- Rule 2: Highly ranked venues attract *many* papers from highly ranked authors.

Let X and Y denote the venue type and author type respectively, and W_{YY} and W_{YX} denote the adjacency matrices for co-author relationships and author-venue relationships in a bibliographic network, according to Rule 1, each author's score is determined by the number of papers and their publication forums,

$$\mathbf{r}_Y(j) = \sum_{i=1}^{m} W_{YX}(j,i)\mathbf{r}_X(i) \tag{1}$$

At the end of each step, $\mathbf{r}_Y(j)$ is normalized by $\mathbf{r}_Y(j) \leftarrow \frac{\mathbf{r}_Y(j)}{\sum_{j'=1}^{n} \mathbf{r}_Y(j')}$.

According to Rule 2, the score of each venue is determined by the quantity and quality of papers in the venue, which is measured by their authors' rank scores,

$$\mathbf{r}_X(i) = \sum_{j=1}^{n} W_{XY}(i,j)\mathbf{r}_Y(j) \tag{2}$$

The score vector is then normalized by $\mathbf{r}_X(i) \leftarrow \frac{\mathbf{r}_X(i)}{\sum_{i'=1}^{m} \mathbf{r}_X(i')}$.

The two formulas will converge to the primary eigenvector of $W_{XY}W_{YX}$ and $W_{YX}W_{XY}$ respectively.

When considering the co-author information, the scoring function can be further refined by a third rule:

- Rule 3: The rank of an author is enhanced if he or she co-authors with many highly ranked authors.

Adding this new rule, we can calculate rank scores for authors by revising Equation (1) as

$$\mathbf{r}_Y(i) = \alpha \sum_{j=1}^{m} W_{YX}(i,j)\mathbf{r}_X(j) + (1-\alpha)\sum_{j=1}^{n} W_{YY}(i,j)\mathbf{r}_Y(j) \tag{3}$$

where parameter $\alpha \in [0,1]$ determines how much weight to put on each factor, which can be assigned based on one's belief or learned by some training dataset.

Similarly, we can prove that $\mathbf{r}_Y$ should be the primary eigenvector of $\alpha W_{YX}W_{XY} + (1-\alpha)W_{YY}$, and $\mathbf{r}_X$ should be the primary eigenvector of $\alpha W_{XY}(I-(1-\alpha)W_{YY})^{-1}W_{YX}$. Since the iterative process is a power method to calculate primary eigenvectors, the rank score will finally converge.

The idea is extended to ranking medical treatments based on medical literature, and an algorithm called MedRank is proposed in [Chen et al(2013)Chen, Li, and Han].

Proximity Ranking

Different from previous ranking methods that either rank objects according to their global importance or find the important objects that are relevant to a query, ranking objects according to their similarity or proximity to a given object is also important.

Note that, proximity ranking does not necessarily return highly visible objects in the network.

SimRank

SimRank is proposed in [Jeh and Widom(2002)] to calculate pairwise similarity between objects in a network based on the link information. The intuition of the similarity model is based on the idea that "two objects are similar if they are related to similar objects." In other words, the similarity between objects can be propagated from pair to pair via links.

For a directed graph $G = (V, E)$, the similarity between two nodes a and b is defined to be 1, if $a = b$, that is, $s(a, b) = 1$ when $a = b$. Otherwise, it is calculated iteratively via the following formula:

$$s(a,b) = \frac{C}{|I(a)||I(b)|} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} s(I_i(a), I_j(b))$$

where C is the damping factor and is set as 0.8 in the paper, $I(a)$ represents the in-neighbors of node a, $|I(a)|$ is the total number of in-neighbors of a, and $I_i(a)$ represents the ith in-neighbor of a.

SimRank can also be applied to bipartite networks, where similarity between one type enhances the quality of the other type alternatively.

It can be shown that SimRank computation on a network G is equivalent to the pairwised random surfer model on a network of G^2. The rank score of a node in G^2 represents the similarity score of a pair of nodes in the original network G. The convergence of the SimRank computation can be guaranteed.

The time complexity of computing SimRank is high, as the similarity score between a pair of objects is dependent on the similarity between every other pair of objects. Different algorithms are proposed to fast computing SimRank, such as [Li et al(2010a)Li, Han, He, Jin, Sun, Yu, and Wu] and [Li et al(2010b)Li, Liu, Xu, Jun, and Du].

PathSim

PathSim [Sun et al(2011)Sun, Han, Yan, Yu, and Wu] is designed to evaluate peer similarity between objects in a heterogeneous information network. Different from previous query-based ranking and similarity measure, PathSim is proposed for (1) evaluating similarity between objects in a *heterogeneous information network*, and (2) evaluating similarity in terms of *peers* between objects.

In heterogeneous information networks, objects can be connected via different types of connections, and similarity with different semantics can be defined using different types of connections. Meta-path, the meta-level connection between objects, is then proposed to systematically capture how objects are connected in a heterogeneous network.

In many scenarios, finding similar objects in networks is to *find similar peers*, such as finding similar authors based on their fields and reputation, finding similar actors based on their movie styles and productivity, and finding similar products based on their functions and popularity. A meta-path-based similarity measure, called *PathSim*, that captures the subtlety of peer similarity, is proposed. The intuition behind it is that two similar peer objects should not only be strongly connected, but also share comparable visibility. Given a symmetric meta-path $\mathcal{P}$, PathSim between two objects x and y of the same type is:

$$s(x,y) = \frac{2 \times |\{p_{x \rightsquigarrow y} : p_{x \rightsquigarrow y} \in \mathcal{P}\}|}{|\{p_{x \rightsquigarrow x} : p_{x \rightsquigarrow x} \in \mathcal{P}\}| + |\{p_{y \rightsquigarrow y} : p_{y \rightsquigarrow y} \in \mathcal{P}\}|}$$

where $p_{x \rightsquigarrow y}$ is a path instance between x and y, $p_{x \rightsquigarrow x}$ is that between x and x, and $p_{y \rightsquigarrow y}$ is that between y and y.

Meta-path-based similarity is a general framework, on which other measures can be defined to evaluate similarity or proximity between objects. For example,

[Shi et al(2012)Shi, Kong, Yu, Xie, and Wu] proposes a proximity measure between different types of objects.

Learning to Rank

Most of the previously discussed ranking methods are un-supervised. However, in may cases, ranking should be different for different datasets and/or for different purposes. Thus, learning is important to select the best parameters for a parameterized ranking method. For example, the previous mentioned PopRank [Nie et al(2005)Nie, Zhang, Wen, and Ma] can automatically learn the best popularity propagation probabilities between object types. Besides PopRank, there are several other recently proposed supervised or semi-supervised ranking methods, as introduced below.

Adaptive PageRank

In [Tsoi et al(2003)Tsoi, Morini, Scarselli, Hagenbuchner, and Maggini], the authors propose to help administrators alter PageRank scores according to their preference by modifying PageRank equations and introducing constraints.

The administrator of a system may want to intervene the PageRank score, such as modify the page scores to some target scores, or establish a predefined ordering on the pages. These constraints can be represented as some linear constraints. At the same time, the administrator wants to find a scoring function that is most similar to the original PageRank scoring function. The problem can then be transformed to a quadratic programming problem with an inequality constraint set. And the parameters can be automatically learned to derive an administrator preferred ranking function.

Learn to Rank Networked Entities (NetRank)

In [Agarwal et al(2006)Agarwal, Chakrabarti, and Aggarwal], the authors propose to parameterize the conductance values between objects, and rank networked entities

based on Markov walks with these parameterized conductance value. The goal is to learn those parameters according to a given preference order among objects.

The conductance value between two objects u and v is defined as the network flow between u and v:

$$p_{uv} = Pr(u \to v) = p_u p(v|u)$$

where p_u is the probability that a random surfer stays at node u, and $p(v|u)$ is the transition probability from u to v.

The conductance value is considered to be parameterized in two ways. First, it can be parameterized according to the hidden communities that the two nodes belong to. Intuitively, edges within the same community have a higher conductance and edges that bridge different communities have a lower conductance. Second, the conductance value can be parameterized according to the edge type that (u, v) belongs to. Intuitively, different types of edges may have different conductance.

Semi-Supervised PageRank

A semi-supervised learning framework, called semi-supervised PageRank, is proposed in [Gao et al(2011)Gao, Liu, Wei, Wang, and Li], which aims at ranking nodes on a very large graph. In the algorithm, the objective function is defined based upon Markov random walk on the graph. The transition probability and the reset probability of the Markov model are defined as parametric models based on the features on both nodes and edges.

For the objective function, the goal is to find a ranking that is as close to the parametric Markov process stationary probability as possible. At the same time, the constraints indicate the guidance from the users, and require that the ranking is as consistent with the user supervision as possible.

It turns out that adaptive PageRank and NetRank are both special cases of the proposed approach.

Similarity Search by Meta-Path Selection

A query-dependent semi-supervised ranking method in heterogeneous information network is proposed in [Yu et al(2012)Yu, Sun, Norick, Mao, and Han], which aims to find entities with high similarity to a given query entity.

Due to the diverse semantic meanings in a heterogeneous information network that contains multi-typed entities and relationships, similarity measurement can be ambiguous without context. A meta-path-based ranking model ensemble is proposed to represent semantic meanings for similarity queries. Users can provide several sample similar objects while issuing the query, and the algorithm will automatically select the best ranking model according to such hints and dispatch the query to the selected ranking model online.

Key Applications

Ranking methods are important for many applications. For example, ranking is critical for search engine systems, either web search or entity search. It can also be used in entity ranking for applications in a particular domain, such as in a bibliographic database or a medical information system. Proximity ranking turns out to be very useful in recommender systems. Identifying the most influential actors in social networks can help viral marketing. Ranking can also be used for spam detection and trustworthy analysis.

Cross-References

Centrality Measures; Data Mining; Eigenvalues, Singular Value Decomposition; Node Ranking in Social Networks; Social Influence Analysis; Social Web Search.

References

[Agarwal et al(2006)Agarwal, Chakrabarti, and Aggarwal] Agarwal A, Chakrabarti S, Aggarwal S (2006) Learning to rank networked entities. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '06, pp 14–23, DOI 10.1145/1150402.1150409, URL http://doi.acm.org/10.1145/1150402.1150409

[Balmin et al(2004)Balmin, Hristidis, and Papakonstantinou] Balmin A, Hristidis V, Papakonstantinou Y (2004) Objectrank: authority-based keyword search in databases. In: Proceedings of the Thirtieth international conference on Very large data bases - Volume 30, VLDB Endowment, VLDB '04, pp 564–575

[Beauchamp(1965)] Beauchamp MA (1965) An improved index of centrality. Behavioral Science 10:161–163

[Brin and Page(1998)] Brin S, Page L (1998) The anatomy of a large-scale hypertextual web search engine. Computer Networks 30(1-7):107–117

[Chen et al(2013)Chen, Li, and Han] Chen L, Li X, Han J (2013) Medrank: Discovering influential medical treatments from literature by information network analysis. In: Proc. 2013 Australasian Database Conf., Adelaide, South Australia, ADC '13

[Freeman(1977)] Freeman LC (1977) A set of measures of centrality based on betweenness. Sociometry 40:35–41

[Freeman(1978)] Freeman LC (1978) Centrality in social networks conceptual clarification. Social Networks 1(3):215–239, DOI 10.1016/0378-8733(78)90021-7

[Gao et al(2011)Gao, Liu, Wei, Wang, and Li] Gao B, Liu TY, Wei W, Wang T, Li H (2011) Semi-supervised ranking on very large graphs with rich metadata. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '11, pp 96–104, DOI 10.1145/2020408.2020430, URL http://doi.acm.org/10.1145/2020408.2020430

[Gyöngyi et al(2004)Gyöngyi, Garcia-Molina, and Pedersen] Gyöngyi Z, Garcia-Molina H, Pedersen J (2004) Combating web spam with trustrank. In: Proceedings of the Thirtieth international conference on Very large data bases - Volume 30, VLDB Endowment, VLDB '04, pp 576–587, URL http://dl.acm.org/citation.cfm?id=1316689.1316740

[Haveliwala(2002)] Haveliwala TH (2002) Topic-sensitive pagerank. In: Proceedings of the 11th international conference on World Wide Web, WWW '02, pp 517–526

[Jeh and Widom(2002)] Jeh G, Widom J (2002) Simrank: a measure of structural-context similarity. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '02, pp 538–543, DOI 10.1145/775047.775126, URL http://doi.acm.org/10.1145/775047.775126

[Jeh and Widom(2003)] Jeh G, Widom J (2003) Scaling personalized web search. In: Proceedings of the 12th international conference on World Wide Web, New York, NY, USA, WWW '03, pp 271–279, DOI 10.1145/775152.775191

[Katz(1953)] Katz L (1953) A new status index derived from sociometric analysis. Psychometrika 18(1):39–43

[Kleinberg(1999)] Kleinberg JM (1999) Authoritative sources in a hyperlinked environment. J ACM 46(5):604–632

[Li et al(2010a)Li, Han, He, Jin, Sun, Yu, and Wu] Li C, Han J, He G, Jin X, Sun Y, Yu Y, Wu T (2010a) Fast computation of simrank for static and dynamic information networks. In: Proceedings of the 13th International Conference on Extending Database Technology, EDBT '10, pp 465–476, DOI 10.1145/1739041.1739098, URL http://doi.acm.org/10.1145/1739041.1739098

[Li et al(2010b)Li, Liu, Xu, Jun, and Du] Li P, Liu H, Xu J, Jun Y, Du HX (2010b) Fast single-pair simrank computation. In: In Proc. of the SIAM Intl. Conf. on Data Mining, SDM '10

[Nie et al(2005)Nie, Zhang, Wen, and Ma] Nie Z, Zhang Y, Wen JR, Ma WY (2005) Object-level ranking: bringing order to web objects. In: Proceedings of the 14th international conference on World Wide Web, WWW '05, pp 567–574, DOI 10.1145/1060745.1060828

[Nieminen(1974)] Nieminen J (1974) On the centrality in a graph. Scandinavian Journal of Psychology 15(1):332–336

[Sabidussi(1966)] Sabidussi G (1966) The centrality index of a graph. Psychomatrika 31:581–603

[Shi et al(2012)Shi, Kong, Yu, Xie, and Wu] Shi C, Kong X, Yu PS, Xie S, Wu B (2012) Relevance search in heterogeneous networks. In: Proceedings of the 15th International Conference on Extending Database Technology, EDBT '12, pp 180–191, DOI 10.1145/2247596.2247618, URL http://doi.acm.org/10.1145/2247596.2247618

[Sun et al(2009a)Sun, Han, Zhao, Yin, Cheng, and Wu] Sun Y, Han J, Zhao P, Yin Z, Cheng H, Wu T (2009a) 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 et al(2009b)Sun, Yu, and Han] Sun Y, Yu Y, Han J (2009b) 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, pp 797–806

[Sun et al(2011)Sun, Han, Yan, Yu, and Wu] Sun Y, Han J, Yan X, Yu PS, Wu T (2011) Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. In: Proc. of 2011 Int. Conf. on Very Large Data Bases, VLDB '11

[Tsoi et al(2003)Tsoi, Morini, Scarselli, Hagenbuchner, and Maggini] Tsoi AC, Morini G, Scarselli F, Hagenbuchner M, Maggini M (2003) Adaptive ranking of web pages. In: Proceedings of the 12th international conference on World Wide Web, WWW '03, pp 356–365, DOI 10.1145/775152.775203, URL http://doi.acm.org/10.1145/775152.775203

[Wasserman and Faust(1994)] Wasserman S, Faust K (1994) Social Network Analysis: Methods and Applications. Cambridge University Press

[Yu et al(2012)Yu, Sun, Norick, Mao, and Han] Yu X, Sun Y, Norick B, Mao T, Han J (2012) User guided entity similarity search using meta-path selection in heterogeneous information networks. In: Proceedings of the 21st ACM international conference on Information and knowledge management, CIKM '12, pp 2025–2029, DOI 10.1145/2396761.2398565