Mining Heterogeneous Information Networks by Exploring the Power of Links

Jiawei Han

Department of Computer Science
University of Illinois at Urbana-Champaign
hanj@cs.uiuc.edu

Abstract. Knowledge is power but for interrelated data, knowledge is often hidden in massive links in heterogeneous information networks. We explore the power of links at mining heterogeneous information networks with several interesting tasks, including link-based object distinction, veracity analysis, multidimensional online analytical processing of heterogeneous information networks, and rank-based clustering. Some recent results of our research that explore the crucial information hidden in links will be introduced, including ((1) Distinct for object distinction analysis, (2) TruthFinder for veracity analysis, (3) Infonet-OLAP for online analytical processing of information networks, and (4) RankClus for integrated ranking-based clustering. We also discuss some of our on-going studies in this direction.

1 Introduction

Social, natural, and information systems usually consist of a large number of interacting, multi-typed components. Examples of such systems include communication and computer systems, the World-Wide Web, biological networks, transportation systems, epidemic networks, criminal rings, and hidden terrorist networks. All the above systems share an important common feature: they are *networked systems*, *i.e.*, individual agents or components interact with a specific set of components, forming large, interconnected, and heterogeneous (*i.e.*, multi-typed) networks. Without loss of generality, we call such interconnected, multi-typed networks or systems as ***heterogeneous information networks***. Clearly, heterogeneous information networks are ubiquitous and form a critical component of modern information infrastructure.

Despite their prevalence in our world, we have only recently recognized the importance of studying information networks as a whole. Hidden in these networks are the answers to important questions. For example, is there a collaborated plot behind a network intrusion, and how can we identify its source in communication networks? How can a company derive a complete view of its products at the retail level from interlinked social communities? These questions are highly relevant to a new class of analytical applications that query and mine massive information networks for pattern and knowledge discovery, data and information integration, veracity analysis and deep understanding of the principles of information networks.

J. Gama et al. (Eds.): DS 2009, LNAI 5808, pp. 13–30, 2009.

Searching for information and knowledge inside networks, particularly large networks with thousands of nodes is a complex and time-consuming task. Unfortunately, the lack of a general analytical and access platform makes sensible navigation and human comprehension virtually impossible in large-scale networks. Fortunately, information networks contains massive nodes and links associated with various kinds of information. Knowledge about such networks is often hidden in massive links in heterogeneous information networks but can be uncovered by the development of sophisticated knowledge discovery mechanisms.

In this paper, we outline some of our recent studies that explore the power of links at mining heterogeneous information networks, including link-based object distinction, veracity analysis, multidimensional online analytical processing of heterogeneous information networks, and rank-based clustering. Such studies show that powerful data mining mechanisms can be used for analysis and exploration of large-scale information networks and systematic development of such network mining methods is an important task in future research.

The remaining of the paper is organized as follows. Section 2 introduces object distinction analysis, Section 3 on veracity analysis, Section 4 on OLAP information networks, and Section 5 on integrated ranking-based clustering. We summarize our study in Section 6.

2 Distinguishing Objects with Identical Names by Information Network Analysis

People retrieve information from different databases on the Web, such as DBLP, Amazon shopping, and AllMusic. One disturbing problem is that different objects may share identical names. For example, there are 72 songs and 3 albums named "Forgotten" in allmusic.com; and there are over 200 papers in DBLP written by at least 14 different Wei Wang's. Users are often unable to distinguish them, because the same object may appear in very different contexts, and there is often limited and noisy information associated with each appearance.

The task of *distinguishing objects with identical names* is called *object distinction* analysis. Given a database and a set of references in it referring to multiple objects with identical names, the task is to split the references into clusters, so that each cluster corresponds to one real object. This task is the opposite of a popular problem called *reference reconciliation* (or *record linkage*, *duplicate detection*), which aims at merging records with different contents referring to the same object, such as two citations referring to the same paper. There have been many approaches developed for record linkage analysis [2], which usually use some efficient techniques [4] to find candidates of duplicate records (*e.g.*, pairs of objects with similar names), and then check duplication for each pair of candidates. Different approaches are used to reconcile each candidate pair, such as probabilistic models of attribute values and textual similarities [2].

Compared with record linkage, objection distinction is a very different problem. First, because the references have identical names, textual similarity is useless. Second, each reference is usually associated with limited information, and thus it is difficult to make good judgement based on it. Third and most importantly, because different references to the same object appear in different contexts, they seldom share common or similar attribute values. Most record linkage approaches are based on the assumption that duplicate records should have equal or similar values, and thus cannot be used on this problem.

Although the references are associated with limited and possibly inconsistent information, the linkages among references and other objects still provide crucial information for grouping references. For example, in a publication database, different references to authors are connected in numerous ways through authors, conferences and citations. References to the same author are often linked in certain ways, such as through their coauthors, coauthors of coauthors, and citations. These linkages provide important information, and a comprehensive analysis on them may likely disclose the identities of objects.

We developed a methodology called Distinct [11] that can distinguish object identities by fusing different types of linkages with differentiating weights, and using a combination of distinct similarity measures to assess the value of each linkage. Because the linkage information is usually sparse and intertwined, Distinct combines two approaches for measuring similarities between records in a relational database: (i) *set resemblance between the neighbor tuples* of two records (the *neighbor tuples* of a record are the tuples linked with it); and (ii) *random walk probability* between two records in the graph of relational data. These two approaches are complementary: one uses the neighborhood information, and the other uses connection strength of linkages. Moreover, since there are many types of linkages among references, each following a join path in the database schema, and different types of linkages have very different semantic meanings and different levels of importance, Distinct uses support vector machines (SVM) to learn a model for weighing different types of linkages. When grouping references, the references to the same object can be merged and considered as a whole. Distinct uses agglomerative hierarchical clustering, which repeatedly merges the most similar pairs of clusters. It combines *average-link* (average similarity between all objects in two clusters) and *collective similarity* (considering each cluster as a single object) to measure the similarity between two clusters, which is less vulnerable to noise.

Distinct uses supervised learning to determine the pertinence of each join path and assign a weight to it. In order to do this, a training set is needed that contains equivalent references as positive examples and distinct references as negative ones. Instead of manually creating a training set which requires much labor and expert knowledge, Distinct constructs the training set automatically, based on the observation that the majority of entities have distinct names in most applications. Take the problem of distinguishing persons as an example. A person's name consists of the first and last names. If a name contains a rather rare first name and a rather rare last name, this name is very likely to be unique. We can find many such names in a database and use them to construct training sets. A pair of references to an object with a unique name can be used as a positive example, and a pair of references to two different objects can be used as a negative example.

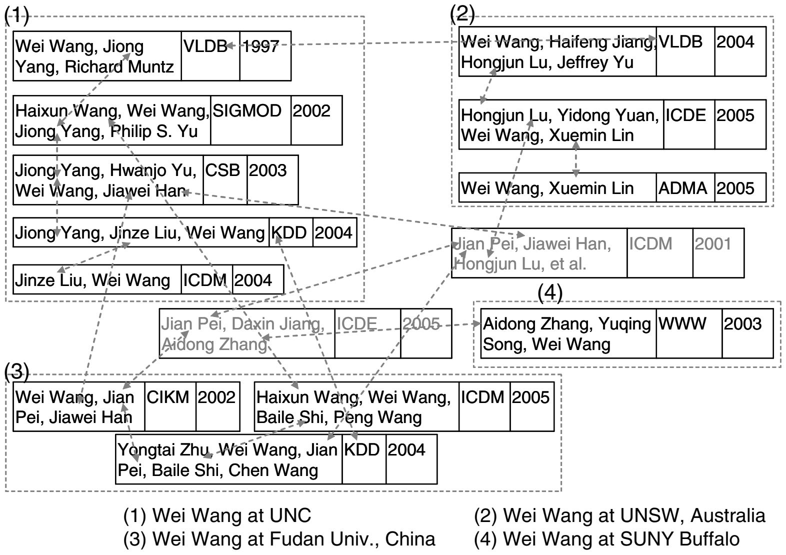

(1)
Wei Wang, Jiong Yang, Richard Muntz | VLDB | 1997
Haixun Wang, Wei Wang, Jiong Yang, Philip S. Yu | SIGMOD | 2002
Jiong Yang, Hwanjo Yu, Wei Wang, Jiawei Han | CSB | 2003
Jiong Yang, Jinze Liu, Wei Wang | KDD | 2004
Jinze Liu, Wei Wang | ICDM | 2004
(2)
Wei Wang, Haifeng Jiang, Hongjun Lu, Jeffrey Yu | VLDB | 2004
Hongjun Lu, Yidong Yuan, Wei Wang, Xuemin Lin | ICDE | 2005
Wei Wang, Xuemin Lin | ADMA | 2005
Jian Pei, Jiawei Han, Hongjun Lu, et al. | ICDM | 2001
(4)
Aidong Zhang, Yuqing Song, Wei Wang | WWW | 2003
Jian Pei, Daxin Jiang, Aidong Zhang | ICDE | 2005
(3)
Wei Wang, Jian Pei, Jiawei Han | CIKM | 2002
Haixun Wang, Wei Wang, Baile Shi, Peng Wang | ICDM | 2005
Yongtai Zhu, Wei Wang, Jian Pei, Baile Shi, Chen Wang | KDD | 2004
(1) Wei Wang at UNC
(2) Wei Wang at UNSW, Australia
(3) Wei Wang at Fudan Univ., China
(4) Wei Wang at SUNY Buffalo

Fig. 1. Papers by four different Wei Wang's

Example 1: Distinguishing people or objects with identical names. There are more than 200 papers in DBLP written by at least 14 different Wei Wang's, each having at least two papers. A mini example is shown in Fig. 1, which contains some papers by four different Wei Wang's and the linkages among them. Users are often unable to distinguish them, because the same person or object may appear in very different contexts, and there is often limited and noisy information associated with each appearance.

We report our empirical study on testing the effectiveness of the proposed approach. Distinct is tested on the DBLP database. First, authors with no more than 2 papers are removed, and there are 127,124 authors left. There are about 616K papers and 1.29M references to authors in *Publish* relation (authorship). In DBLP we focus on distinguishing references to authors with identical names.

We first build a training set using the method illustrated above, which contains 1000 positive and 1000 negative examples. Then SVM with linear kernel is applied. We measure the performance of Distinct by precision, recall, and f-measure. Suppose the standard set of clusters is C^*, and the set of clusters by Distinct is C. Let TP (true positive) be the number of pairs of references that are in the same cluster in both C^* and C. Let FP (false positive) be the number of pairs of references in the same cluster in C but not in C^*, and FN (false negative) be the number of pairs of references in the same cluster in C^* but not in C.

$$precision = \frac{TP}{TP+FP}, \qquad recall = \frac{TP}{TP+FN}.$$

f-measure is the harmonic mean of precision and recall.

Table 1. Names corresponding to multiple authors

Name	#author	#ref	Name	#author	#ref
Hui Fang	3	9	Bing Liu	6	89
Ajay Gupta	4	16	Jim Smith	3	19
Joseph Hellerstein	2	151	Lei Wang	13	55
Rakesh Kumar	2	36	Wei Wang	14	141
Michael Wagner	5	29	Bin Yu	5	44

Table 2. Accuracy for distinguishing references

Name	precision	recall	f-measure
Hui Fang	1.0	1.0	1.0
Ajay Gupta	1.0	1.0	1.0
Joseph Hellerstein	1.0	0.810	0.895
Rakesh Kumar	1.0	1.0	1.0
Michael Wagner	1.0	0.395	0.566
Bing Liu	1.0	0.825	0.904
Jim Smith	0.888	0.926	0.906
Lei Wang	0.920	0.932	0.926
Wei Wang	0.855	0.814	0.834
Bin Yu	1.0	0.658	0.794
average	0.966	0.836	0.883

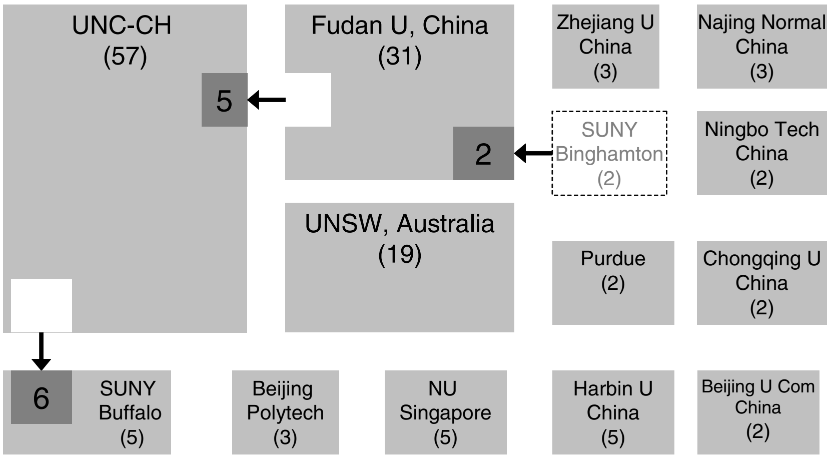


Fig. 2. Groups of references of "Wei Wang"

We test Distinct on real names in DBLP that correspond to multiple authors. 10 such names are shown in Table 1, together with the number of authors and number of references. For each name, we manually divide the references into groups according to the authors' identities, which are determined by the authors' home pages or affiliations shown on the papers.

We use Distinct to distinguish references to each name, with min-sim set to 0.0005. Table 2 shows the performance of Distinct for each name. In general, Distinct successfully group references with high accuracy. There is no false positive in 7 out of 10 cases, and the average recall is 83.6%. In some cases references to one author are divided into multiple groups. For example, 18 references to "Michael Wagner" in Australia are divided into two groups, which leads to low recall.

We visualize the results about "Wei Wang" in Fig. 2. References corresponding to each author are shown in a gray box, together with his/her current affiliation and number of references. The arrows and small blocks indicate the mistakes made by Distinct. It can be seen that in general Distinct does a very good job in distinguishing references, although it makes some mistakes because of the linkages between references to different authors.

3 Truth Discovery with Multiple Conflicting Information Providers

Information networks nowadays are fed with tremendous amounts of data from numerous information sources. These sources may provide *conflicting* information about the *same* entity, and pieces of information on the web could be already outdated when being read. This problem will only go worse since more information will be available on the web, and such conflicting information could become norm instead of exception. Therefore, it is necessary to provide trustable analysis of the truthfulness of information from multiple information providers and automatically identify the correct information. Such truth validation analysis is called veracity analysis.

Example 2. Veracity analysis on the authors of books provided by online bookstores. People retrieve all kinds of information from the web everyday. When shopping online, people find product specifications and sales information from various web sites like Amazon.com or ShopZilla.com. When looking for interesting DVDs, they get information and read movie reviews on web sites such as NetFlix.com or IMDB.com. Almost for any kind of products, there exist hundred of sale agents and information providers, if not more. *Is the information provided on the Web always trustable?* Unfortunately, the answer is negative. There is no guarantee for the correctness of information on the web. Even worse, different web information providers often present conflicting information. For example, we have found that there are multiple versions for the sets of authors of the same book, titled "Rapid Contextual Design" (ISBN: 0123540518), provided by different online bookstores, as shown in Table 3. From the image of the book cover, one can see that *A1 Books* provides the most accurate information. On the other hand, the information from *Powell's books* is incomplete, and that from *Lakeside books* is incorrect.

To analyze such a problem, the data sets can be viewed as a heterogeneous information network, consisting of three types of objects: (i) information providers (*e.g.*, online bookstores), (ii) objects (*e.g.*, books), and (iii) stated facts about

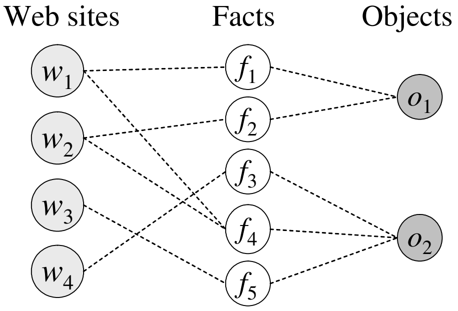


Fig. 3. A network snapshot of information providers, objects, and stated facts

Table 3. Conflicting information about book authors

Web site	*Authors*
A1 Books	Karen Holtzblatt, Jessamyn Burns Wendell, Shelley Wood
Powell's books	Holtzblatt, Karen
Cornwall books	Holtzblatt-Karen, Wendell-Jessamyn Burns, Wood
Mellon's books	Wendell, Jessamyn
Lakeside books	WENDELL, JESSAMYNHOLTZBLATT, KARENWOOD, SHELLEY
Blackwell online	Wendell, Jessamyn, Holtzblatt, Karen, Wood, Shelley
Barnes & Noble	Karen Holtzblatt, Jessamyn Wendell, Shelley Wood

the object (*i.e.*, claimed set of authors). One such mini-example is shown in Figure 3, which contains five stated facts about two objects provided by four web sites. Each web site provides at most one fact for an object.

There have been many studies on ranking web pages according to authority (or popularity) based on hyperlinks, such as Authority-Hub analysis [7], and PageRank [8]. However, top-ranked web sites may not be the most accurate ones. For example, according to our experiments the bookstores ranked on the very top ones by Google (which are *Barnes & Noble* and *Powell's books*) contain more errors on book author information than some small bookstores (*e.g.*, *A1 Books*) that provide more accurate information.

The problem of discovery of trustable information based on those provided multiple information providers is called the *veracity analysis* problem. It can be stated as follows: *Given a large amount of conflicting information about many objects, which is provided by multiple web sites (or other types of information providers), veracity analysis is to discover the true fact about each object.* Here the word *"fact"* is used to represent something that is claimed as a fact by some web site, and such a fact can be either true or false. Notice that here we only investigate the facts that are either the properties of objects (*e.g.*, weights of laptop computers), or the relationships between two objects (*e.g.*, authors of books).

Our solution is a TruthFinder framework [12], that finds confidently the true facts and trustworthy web sites. The method examines the relationships between information providers and the information they provided, with the following two major heuristics: (1) *an assertion that several information providers agree on is usually more trustable than that only one provider suggests*; and (2) *an information provider is* trustworthy *if it provides many pieces of true information, and a piece of information is likely to be true if it is provided by many trustworthy web sites.* The method links three types of information: (i) the information providers, (ii) stated facts on different entities, and (iii) the corresponding entities, into a heterogeneous information network, and performs an in-depth information network analysis. It starts with no bias on a particular piece of information, but uses the above heuristics to derive initial weights on the trustworthiness of the stated facts and information providers. Then it consolidates the trustworthiness by an iterative enhancement process with weight-propagation and consolidation across this information network. The process is similar to the page ranking process proposed in the PageRank and HITS algorithms [1,7] but the weight to be iteratively revised is the trustworthiness probability rather than authority score. We tested TruthFinder on the book author information provided by book sellers on the Web. The method successfully finds facts about who are the true set of authors and who are the trustable information providers.

Table 4 shows the accuracy of TruthFinder in comparison with that of Voting and Barnes & Noble on determining the authors for 100 randomly selected books, where *Voting* is a simple voting among multiple information providers, *i.e.*, considering the fact provided by a majority of web sites as the true fact. The result shows that TruthFinder achieves high accuracy at finding trustable information.

The TruthFinder methodology, though interesting, has two disadvantages: (1) it takes only one version of truth, and does not recognize there could be *multiple versions of truth*: the judgement of an event or an opinion could be rather different from people to people, *e.g.*, the view on a candidate in an election could be rather different but could be clustered into two to three views; and (2) it does not consider *truth may have timeliness*: *e.g.*, a player could win first but then lose, and depending on when the news was delivered, you may get rather different results.

Our on-going research is to overcome these two limitations and perform veracity analysis in information networks. First, we assume that there are *multiple*

Table 4. Performance comparison on a set of books among three methods: Voting, TruthFinder, and Barnes & Noble

Type of error	Voting	TruthFinder	Barnes & Noble
correct	71	85	64
miss author(s)	12	2	4
incomplete names	18	5	6
wrong first/middle names	1	1	3
has redundant names	0	2	23
add incorrect names	1	5	5
no information	0	0	2

versions of truth, each associated with one cluster of information providers. Second, we adopt the model of *timeliness of truth*, *i.e.*, truth may change with time in a dynamic, interconnected world. Moreover, we take higher priority on the most recent claim as the *up-to-date truth*. However, there are still information providers that deliver false information. Based on these assumptions, we are building a multi-version truth model using an integrated link analysis, information aggregation, and clustering, and consolidate the trustworthiness by an iterative enhancement process with weight-propagation across this information network.

4 On-Line Analytical Processing (OLAP) of Heterogeneous Information Networks

In relational database and data warehouse systems, On-Line Analytical Processing (OLAP) has become a powerful component in multidimensional data analysis. By constructing data cubes [5] over the underlying data and providing easy navigation, OLAP gives users the capability of interactive, multi-dimensional and multi-level analysis over a vast amount of data, with a wide variety of views. Certainly, for more complicated information network data, such capability is greatly needed. "*Can we OLAP information networks?*" In our recent study [3], we address this problem and aim for developing effective and scalable methods for on-line multidimensional analysis of heterogeneous information networks.

There are four major research challenges in Information Network OLAP (*i.e.*, *Infonet-OLAP*): (1) multi-dimensionality: each node/link in a network contains valuable, multi-typed, and multidimensional information, such as multi-level concepts/abstraction, textual contents, spatiotemporal information, and other properties; (2) scalability: information networks are often very large with millions of nodes and edges; (3) flexibility: for the same set of data, different users/applications may like to view and analyze the network dramatically differently, which may lead to the efficient formation and exploration of very different information networks; and (4) quality: information networks may contain noisy, inconsistent, and inter-dependent data.

The concept of Infonet-OLAP can be briefly introduced using *bibliographic networks* extracted from the DBLP website (http://dblp.uni-trier.de). DBLP is an online bibliographic database for computer science conference proceedings and journals, indexing more than one million publications and more than 10,000 proceedings and journals. Each entry at DBLP contains (at least) the following pieces of information, $P : (\langle A_1, \ldots, A_k\rangle, T, V, Y)$, indicating that paper P is coauthored by k researchers, $A_1, \ldots, A_k$, with title T, and published at venue V in year Y. This entry consists of multidimensional information: *Authors, Title, Venue,* and *Time*, which can be viewed at multiple levels of abstraction and in a multi-dimensional space. For example, an author could be a junior author vs. a senior one; a prolific one vs. a nonproductive one; and a venue can be viewed similarly, such as a database venue vs. an AI one, a long-history one vs. a new one, and a highly reputed one vs. a low quality one. One also could apply a concept hierarchy to grouping papers according to their contents. Entries in such a database form a gigantic information network. Moreover, by linking with ACM Digital Library, Citeseer, and Google Scholar, citation information can be integrated as well.

This bibliographic network contains huge amounts of rich, heterogeneous, multidimensional, and temporal information. Users may like to view and analyze the network from different angles, which may lead to the *"formation" of different network views*, *e.g.*, coauthor network, conference network, citation network, and author-theme network. Moreover, some may want to examine a network including only the *selected* research themes (*e.g.*, data mining); whereas others may want to analyze hot topics in highly regarded (*i.e.*, highly *ranked*) conferences. Furthermore, some may like to *roll-up* authors to find how junior and senior researchers collaborate; whereas others would like to see the *evolution* of a theme. Clearly, different applications may require the extraction and analysis of multiple, different information networks involving time-variant, multidimensional, heterogeneous entities. A single, homogeneous, static network view cannot satisfy such flexible needs. The focus of Infonet-OLAP is to provide a general OLAP platform, rather than developing yet another specific network mining algorithm or theory.

Let us examine dimensions at first. Actually, there are two types of dimensions in Infonet-OLAP. The first one, called *informational dimension* (or *Info-Dim*, for short), utilizes informational attributes attached at the whole snapshot level. Suppose the following concept hierarchies are associated with *venue* and *time*:

- *venue*: *conference* → *area* → *all*,
- *time*: *year* → *decade* → *all*;

The role of these two dimensions is to organize snapshots into groups based on different perspectives, *e.g.*, *(db-conf, 2004)* and *(sigmod, all-years)*, where each of these groups corresponds to a "cell" in the OLAP terminology. They control what snapshots are to be looked at, without touching the inside of any single snapshot.

Figure 4 shows such an example where the roll-up is first performed on the dimension *venue* to *db-conf* in individual year (*i.e.*., merging the graphs of $\langle SIGMOD, 2004\rangle$, ..., $\langle VLDB, 2004\rangle$ to ⟨*db-conf*, 2004⟩) and then on the dimension *time* to ⟨*db-conf*, *all-years*⟩.

Second, for the subset of snapshots within each cell, one can summarize them by computing a measure as we did in traditional OLAP. In the Infonet-OLAP context, this gives rise to an *aggregated graph*. For example, a summary network displaying total collaboration frequencies can be achieved by overlaying all snapshots together and summing up the respective edge weights, so that each link now indicates two persons' collaboration activities in the DB conferences of 2004 or during the whole history of SIGMOD.

The second type of dimension, called *topological dimension* (or *Topo-Dim* for short), is provided to operate on nodes and edges within individual

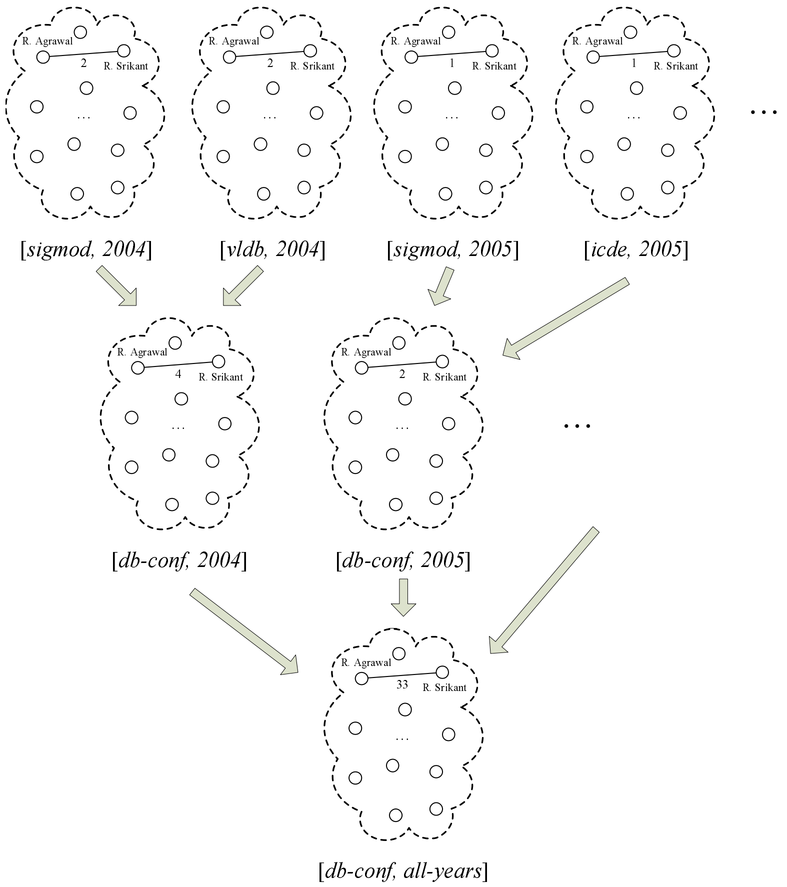


Fig. 4. An I-OLAP Scenario on the DBLP network

networks. Take the DBLP database for instance, suppose the following concept hierarchy

- *authorID*: *individual* → *department* → *institution* → *all*

is associated with the node attribute *authorID*, then it can be used to group authors from the same institution into a "generalized" node, and a new network thus formed will depict interactions among these groups as a whole, which summarizes the original network and hides specific details.

Figure 5 shows such an example on DBLP where the roll-up is performed on the dimension *authorID* to the level *institution*, which merge all persons in the same institution as one node and constructing a new summary graph at the institution level. In the "generalized network", an edge between Stanford and University of Wisconsin will aggregate all collaboration frequencies incurred between Stanford authors and Wisconsin authors. Notice that a roll-up from the individual level to the institution level is achieved by consolidating multiple nodes into one, which shrinks the original network.

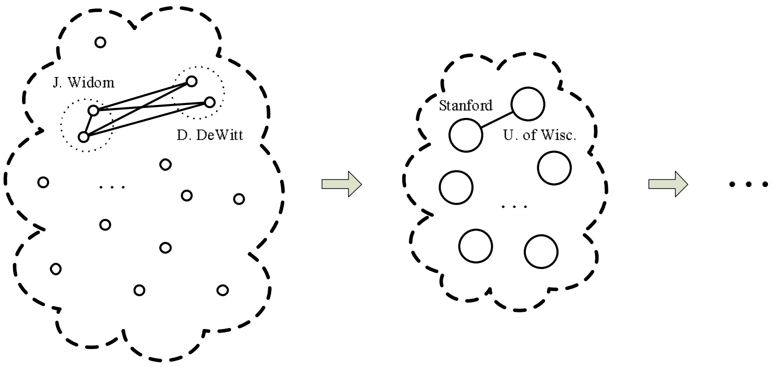


Fig. 5. A T-OLAP Scenario on the DBLP network

The OLAP semantics accomplished through Info-Dims and Topo-Dims are rather different. The first is called *informational OLAP* (abbr. *I-OLAP*), and the second *topological OLAP* (abbr. *T-OLAP*). For roll-up in I-OLAP, the characterizing feature is that, snapshots are just different observations of the same underlying network, and thus when they are all grouped into one cell in the cube, it is like *overlaying* multiple pieces of information, *without changing* the objects whose interactions are being looked at.

For roll-up in T-OLAP, we are no longer grouping snapshots, and the reorganization switches to happen inside individual networks. Here, *merging* is performed internally which "zooms out" the user's focus to a "generalized" set of objects, and a new information network formed by such *shrinking* might greatly alter the original network's topological structure.

As to measures, in traditional OLAP, a measure is calculated by aggregating all the data tuples whose dimensions are of the same values (based on concept hierarchies, such values could range from the finest un-generalized ones to "all/*", which form a multi-level cuboid lattice); casting this to our scenario here:

First, in infonetOLAP, the aggregation of graphs should also take the form of a graph, *i.e.*, an *aggregated graph.* In this sense, graph can be viewed as a special existence, which plays a dual role: as a data source and as an aggregated measure. Of course, other measures that are not graphs, such as node count, average degree, diameter, *etc.*, can also be calculated; however, we do not explicitly include such non-graph measures in our model, but instead treat them as derived from corresponding graph measures.

Second, due to the different semantics of I-OLAP and T-OLAP, aggregating data with identical Info-Dim values groups information among the snapshots, whereas aggregating data with identical Topo-Dim values groups topological elements inside individual networks. As a result, we will give a separate measure definition for each case in below.

A general framework of Infonet-OLAP is presented in [3], however, a systematic study on flexible and efficient implementation of different OALP operations on information networks is still an interesting issue for future research.

5 RankClus: Integrated Ranking-Based Clustering

Besides applying some typical knowledge discovery functions, some process may involve induction on the entire or a substantial portion of the information networks. For example, in order to partition an interconnected, heterogeneous information network into a set of clusters and rank the nodes in each cluster, we have recently develop a RankClus framework [9], which integrates clustering and ranking together to effectively cluster information networks into multiple groups and rank nodes in each group based on certain nice properties (such as authority). Interestingly, such clustering-ranking can be performed based on the links only, even without using the citation information nor the keyword/text information contained in the conferences and/or publication titles. We outline the method in more detail below.

In DBLP, by examining authors, research papers, and conferences, one can group conferences in the same fields together to form conference clusters, group authors based on their publication venues into author clusters and in the meantime rank authors and conferences based on their corresponding authorities. Such an integrated clustering and ranking framework would be an ideal feed for Infonet-OLAP.

Ranking and clustering can provide overall views on information network data, and each has been a hot topic by itself. However, in a large information network, ranking objects globally without considering the clusters they belong to often leads to dumb results, *e.g.*, ranking authors/papers in database/data_mining (DB/DM) and hardware/computer_architcture (HW/CA) conferences together, as shown in Figure 6(a), may not make much sense. Similarly, presenting a huge cluster with numerous entities without distinction is dull as well. Therefore, we propose an integrated approach, called RankClus, to perform ranking and clustering together. Figure 6(b) shows that RankClus can generate meaningful clusters and rankings, even at the expert level, without referring to any citation or content information in DBLP. These clustering and ranking results could be potential input to Infonet-OLAP for effective data mining. RankClus also shows the importance of developing a general Infonet-OLAP framework that allows users to explore interactively with underlying networks for subtle knowledge discovery. An isolated clustering or ranking algorithm is often not enough for such tasks.

According to RankClus, we view the DBLP network as a *bi-type information network* with *conferences* as one type and *authors* as the other. Given two types of object sets X and Y, where $X = \{x_1, x_2, \ldots, x_m\}$, and $Y = \{y_1, y_2, \ldots, y_n\}$, graph $G = \langle V, E\rangle$ is called a bi-type information network on types X and Y, if $V(G) = X \cup Y$ and $E(G) = \{\langle o_i, o_j\rangle\}$, where $o_i, o_j \in X \cup Y$.

Let $W_{(m+n)\times(m+n)} = \{w_{o_i o_j}\}$ be the adjacency matrix of links, where $w_{o_i o_j}$ equals to the weight of link $\langle o_i, o_j\rangle$, which is the number of observations of the link, we thus use $G = \langle\{X \cup Y\}, W\rangle$ to denote this bi-type information network. In the following, we use X and Y denoting both the object set and their type name. For convenience, we decompose the link matrix into four blocks, $W = \begin{pmatrix} W_{XX} & W_{XY} \\ W_{YX} & W_{YY} \end{pmatrix}$, each denoting a sub-network of objects between types of the subscripts.

Rank	Conf.	Rank	Authors
1	DAC	1	Alberto L. Sangiovanni-Vincentelli
2	ICCAD	2	Robert K. Brayton
3	DATE	3	Massoud Pedram
4	ISLPED	4	Miodrag Potkonjak
5	VTS	5	Andrew B. Kahng
6	CODES	6	Kwang-Ting Cheng
7	ISCA	7	Lawrence T. Pileggi
8	VLDB	8	David Blaauw
9	SIGMOD	9	Jason Cong
10	ICDE	10	D. F. Wong

(a) Top-10 ranked conf's/authors in the mixed conf. set

Rank	Conf.	Rank	Authors
1	VLDB	1	H. V. Jagadish
2	SIGMOD	2	Surajit Chaudhuri
3	ICDE	3	Divesh Srivastava
4	PODS	4	Michael Stonebraker
5	KDD	5	Hector Garcia-Molina
6	CIKM	6	Jeffrey F. Naughton
7	ICDM	7	David J. DeWitt
8	PAKDD	8	Jiawei Han
9	ICDT	9	Rakesh Agrawal
10	PKDD	10	Raghu Ramakrishnan

(b) Top-10 ranked conf's/authors in DB/DM set

Fig. 6. RankClus Performs High-Quality Clustering and Ranking Together

Given a bi-type network $G = \langle \{X \cup Y\}, W \rangle$, if a function $f : G \rightarrow (\boldsymbol{r}_X, \boldsymbol{r}_Y)$ gives rank score for each object in type X or type Y, where $\sum_{x \in X} \boldsymbol{r}_X(x) = 1$ and $\sum_{y \in Y} \boldsymbol{r}_Y(y) = 1$. We call f a *ranking function* on network G. Similarly, we can define *conditional rank* and *within-cluster rank*, as follows. Given target type X, and a cluster $X' \subseteq X$, sub-network $G' = \langle \{X' \cup Y\}, W' \rangle$ is defined as a vertex induced graph of G by a vertex subset $X' \cup Y$. *Conditional rank* over Y, denoted as $\boldsymbol{r}_{Y|X'}$, and *within-cluster rank* over X', denoted as $\boldsymbol{r}_{X'|X'}$, are defined by the ranking function f on the sub-network G': $(\boldsymbol{r}_{X'|X'}, \boldsymbol{r}_{Y|X'}) = f(G')$. Conditional rank over X, denoted as $\boldsymbol{r}_{X|X'}$, is defined as the propagation score of $\boldsymbol{r}_{Y|X'}$ over network G:

$$\boldsymbol{r}_{X|X'}(x) = \frac{\sum_{j=1}^{n} W_{XY}(x,j)\boldsymbol{r}_{Y|X'}(j)}{\sum_{i=1}^{m} \sum_{j=1}^{n} W_{XY}(i,j)\boldsymbol{r}_{Y|X'}(j)}. \quad (1)$$

Given a bi-type network $G = \langle \{X \cup Y\}, W \rangle$, the target type X, and K, a specified number of clusters, our goal is to generate K clusters $\{X_k\}, k = 1, 2, \ldots, K$ on X, as well as the within-cluster rank for type X and conditional rank for type Y to each cluster, *i.e.*, $\boldsymbol{r}_{X|X_k}$ and $\boldsymbol{r}_{Y|X_k}, k = 1, 2, \ldots, K$.

In order to generate effective clusters based on authoritative ranking, we need to specify a few rules based on prior knowledge. For example, for the DBLP network, we have the following three empirical rules that will affect our clustering results.

Rule 1: Highly ranked authors publish many papers in highly ranked conferences.

According to this rule, each author's score is determined by the number of papers and their publication forums, *i.e.*, $\boldsymbol{r}_Y(j) = \sum_{i=1}^{m} W_{YX}(j,i)\boldsymbol{r}_X(i)$. This implies when author j publishes more papers, there are more nonzero and high weighted $W_{YX}(j,i)$, and when the author publishes papers in a higher ranked conference i, which means a higher $\boldsymbol{r}_X(i)$, the score of author j will be higher. At the end of each step, $\boldsymbol{r}_Y(j)$ is normalized by $\boldsymbol{r}_Y(j) \leftarrow \frac{\boldsymbol{r}_Y(j)}{\sum_{j'=1}^{n}\boldsymbol{r}_Y(j')}$.

Rule 2: Highly ranked conferences attract many papers from many highly ranked authors.

According to this rule, the score of each conference is determined by the quantity and quality of papers in the conference, which is measured by their authors' ranking scores, *i.e.*, $\boldsymbol{r}_X(i) = \sum_{j=1}^{n} W_{XY}(i,j)\boldsymbol{r}_Y(j)$. This implies when there are more papers appearing in conference i, there are more non-zero and high weighted $W_{XY}(i,j)$; and if the papers are published by higher ranked author j, the rank score for j, which is $\boldsymbol{r}_Y(j)$, is higher, and thus the higher score the conference i will get.The score vector is then normalized by $\boldsymbol{r}_X(i) \leftarrow \frac{\boldsymbol{r}_X(i)}{\sum_{i'=1}^{m}\boldsymbol{r}_X(i')}$. Notice that the normalization will not change the ranking position of an object, but it gives a relative importance score to each object. When considering the co-author information, the scoring function can be further refined by the third heuristic:

Rule 3: Highly ranked authors usually co-author with many authors or many highly ranked authors.

Based on this rule, we can revise the above two equations as

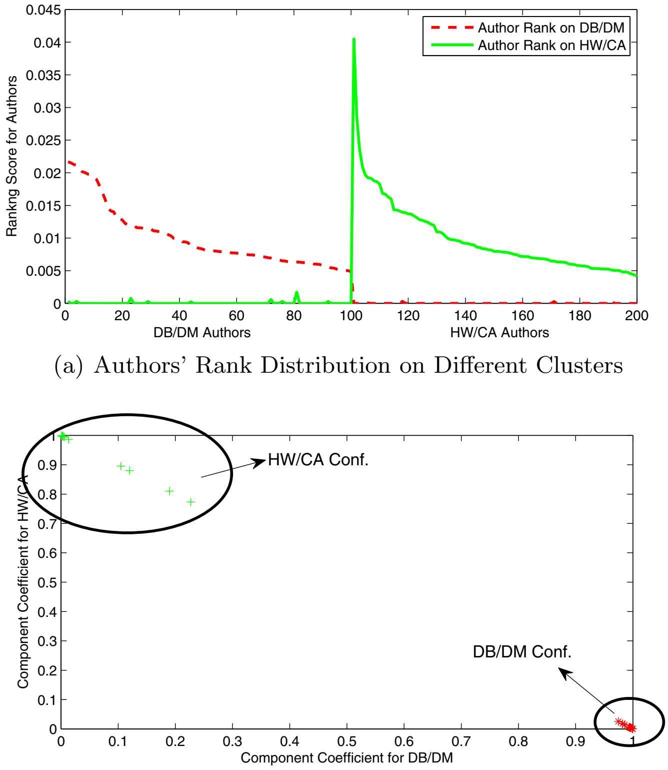
$$\boldsymbol{r}_Y(i) = \alpha\sum_{j=1}^{m} W_{YX}(i,j)\boldsymbol{r}_X(j) + (1-\alpha)\sum_{j=1}^{n} W_{YY}(i,j)\boldsymbol{r}_Y(j). \tag{2}$$

where $\alpha \in [0,1]$ is a weighting coefficient, which could be learned by a training set.

Notice such rules should be worked out by experts based on their research experience in a specific field. For example, in PubMed domain, people may emphasize more on journal publications than conference ones. Also, subtlety exists in certain rules. For example, for Rule 1, a conference/journal will not be reputed if it attracts papers *only* from a tiny group of "prolific" authors because this tiny group may set up its own venue and publish many papers *only* there. They could be "prolific" but few other prolific authors would like to join. Thus neither the venue nor the authors can be "reputed" according to the rule.

Traditional graph clustering methods usually tackle one network only during a clustering process. In contrast, with Rule 3, RankClus is able to combine two information networks, *i.e.*, author-conference bipartite network and co-author network, together for better clustering and ranking.

Putting these rules together, RankClus first randomly generates an initial clustering for target objects. It then performs the following three steps repeatedly until the clustering does not change significantly: (1) rank each cluster, by calculating conditional rank for types Y and X and within-cluster rank for type



(a) Authors' Rank Distribution on Different Clusters

(b) Two Component Coefficients of Conferences

Fig. 7. Effectiveness of RankClus at Clustering and Ranking of the DBLP Data

X; (2) estimate coefficients in the mixture model component; and (3) adjust membership in clusters.

Our implementation and testing of the RankClus method have demonstrated the high promise of this approach. For processing efficiency, it is an order of magnitude faster than the well-cited SimRank algorithm [6] because SimRank has to calculate the pairwise similarity between every two objects of the same type, whereas RankClus uses conditional ranking as the measure of clusters, and only need to calculate the distances between each object and the cluster center. For effectiveness, Figure 7(a) shows that conditional rank can be used as cluster features: DB/DM authors rank high with respect to DB/DM conferences, but rank extremely low with respect to HW/CA conferences. Figure 7(b) is the scatter plot for each conference's two component coefficients. The figure shows that component coefficients can be effectively used as object attributes: the two kinds of conferences are separated clearly under the new attributes.

Our on-going work is to extend the RankClus framework from the following three perspectives: (1) The patterns described above are mined without using citation or title information. By adding title, citation, and/or abstract information, one can uncover the network structure at rather deep levels of research topic hierarchies. These hierarchies will provide additional dimensions to Infonet-OLAP and enhance discovery-driven OLAP. (2) Infonet-OLAP also introduces new challenges to RankClus. There could be multiple information networks available. For instance, we could add author-article network and citation network into the above example for better ranking. In our current implementation, RankClus can only integrate two networks together. Our recent study extends this framework into NetClus star-network schema [10], and subsequent studies will accommodate multiple networks, multiple hierarchies and constraints in the RankClus framework. (3) To facilitate easy exploration (drill-down and roll-up) of clustering results in information networks with different granularity, it is preferred to have clustering consistent across different levels of abstraction. The principles of such clustering design will be examined for a collection of multi-resolution information networks.

6 Conclusions

In this paper, we have outlined our new research progress on mining knowledge from heterogeneous information networks. We show that heterogeneous information networks are ubiquitous and with broad applications. Moreover, knowledge is often hidden in massive links in heterogeneous information networks, and thus it is necessary to perform a systematic study on how ro explore the power of links at mining heterogeneous information networks. We presented in several interesting link-mining tasks, including link-based object distinction, veracity analysis, multidimensional online analytical processing of heterogeneous information networks, and rank-based clustering. We show some interesting results of our recent research that explore the crucial information hidden in links will be introduced, including four new methods: (1) Distinct for object distinction analysis, (2) TruthFinder for veracity analysis, (3) Infonet-OLAP for online analytical processing of information networks, and (4) RankClus for integrated ranking-based clustering. We also show that mining heterogeneous information networks is an exciting research frontier and there is much space to be explored in future research.

Acknowledgement

The work was supported in part by the U.S. National Science Foundation grants IIS-08-42769 and BDI-05-15813, NASA grant NNX08AC35A, and the Air Force Office of Scientific Research MURI award FA9550-08-1-0265. Any opinions, findings, and conclusions expressed here are those of the authors and do not necessarily reflect the views of the funding agencies. This paper is a summarization of several pieces of work in our research. The author would like to express his sincere thanks to all the collaborators, especially Chen Chen, Yizhou Sun, Xifeng Yan, Xiaoxin Yin, Philip Yu, and Feida Zhu.

References

1. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. In: Proc. 7th Int. World Wide Web Conf (WWW 1998), Brisbane, Australia, April 1998, pp. 107–117 (1998)
2. Chaudhuri, S., Ganjam, K., Ganti, V., Motwani, R.: Robust and efficient fuzzy match for online data cleaning. In: Proc. 2003 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD 2003), San Diego, CA (June 2003)
3. Chen, C., Yan, X., Zhu, F., Han, J., Yu, P.S.: Graph OLAP: Towards online analytical processing on graphs. In: Proc. 2008 Int. Conf. on Data Mining (ICDM 2008), Pisa, Italy (December 2008)
4. Gravano, L., Ipeirotis, P., Jagadish, H., Koudas, N., Muthukrishnan, S., Srivastava, D.: Approximate string joins in a database (almost) for free. In: Proc. 2001 Int. Conf. Very Large Data Bases (VLDB 2001), Rome, Italy, September 2001, pp. 491–500 (2001)
5. Gray, J., Chaudhuri, S., Bosworth, A., Layman, A., Reichart, D., Venkatrao, M., Pellow, F., Pirahesh, H.: Data cube: A relational aggregation operator generalizing group-by, cross-tab and sub-totals. Data Mining and Knowledge Discovery 1, 29–54 (1997)
6. Jeh, G., Widom, J.: SimRank: a measure of structural-context similarity. In: Proc. 2002 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD 2002), Edmonton, Canada, July 2002, pp. 538–543 (2002)
7. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. J. ACM 46, 604–632 (1999)
8. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Technical Report, Computer Science Dept., Stanford University (1998)
9. Sun, Y., Han, J., Zhao, P., Yin, Z., Cheng, H., Wu, T.: RankClus: Integrating clustering with ranking for heterogeneous information network analysis. In: Proc. 2009 Int. Conf. on Extending Data Base Technology (EDBT 2009), Saint-Petersburg, Russia (March 2009)
10. Sun, Y., Yu, Y., Han, J.: Ranking-based clustering of heterogeneous information networks with star network schema. In: Proc. 2009 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD 2009), Paris, France (June 2009)
11. Yin, X., Han, J., Yu, P.S.: Object distinction: Distinguishing objects with identical names by link analysis. In: Proc. 2007 Int. Conf. Data Engineering (ICDE 2007), Istanbul, Turkey (April 2007)
12. Yin, X., Han, J., Yu, P.S.: Truth discovery with multiple conflicting information providers on the web. IEEE Trans. Knowledge and Data Eng. 20, 796–808 (2008)