Relevance Measure in Large-Scale Heterogeneous Networks

Xiaofeng Meng, Chuan Shi, Yitong Li, Lei Zhang, and Bin Wu

Beijing University of Posts and Telecommunications, Beijing, China 100876

Abstract. Recently, there is a surge of heterogeneous information network analysis, where network includes multiple types of objects or links. Many data mining tasks have been studied on it, among which similarity measure is a basic and important function. Several similarity measures have been proposed in heterogeneous information network. However, they suffer from high computation and memory demand. In this paper, we propose a novel measure, called AvgSim, which can measure similarity of same or different-typed object pairs in a uniform framework and has some good properties. **AvgSim** value of two objects is evaluated through two random walk processes along the given meta-path and the reverse meta-path, respectively. In addition, we implement AvgSim using MapReduce parallel model in order to enable the application in large-scale networks. Experiments on real data sets verify the effectiveness and efficiency of AvgSim.

Keywords: Heterogeneous information network, Similarity search, Random walk, MapReduce.

1 Introduction

In recent years, heterogeneous information network analysis has become a hot research topic in data mining field. Different from widely used homogeneous networks which include only same-typed objects or links, Heterogeneous Information Network (HIN) organizes the networked data as a network including different-typed objects and links. For example, in the case of bibliographic network, the object types include authors, papers, venues and links between objects correspond to different relations, such as write relation between authors and papers, and citation relation between papers. Fig.1(a) and Fig.1(b) shows two bibliographic information network schemas which are ACM dataset and DBLP dataset. Combination of different-typed objects and links results in more comprehensive structure information and rich semantics information. Thus, heterogeneous information network analysis will mine more interesting patterns.

Many data mining tasks have been exploited in heterogeneous information network, such as clustering [1], classification [2]. Among these data mining tasks, similarity measure is a basic and important function, which evaluate the similarity of object pairs on networks. Although similarity measure on homogeneous

L. Chen et al. (Eds.): APWeb 2014, LNCS 8709, pp. 636–643, 2014.

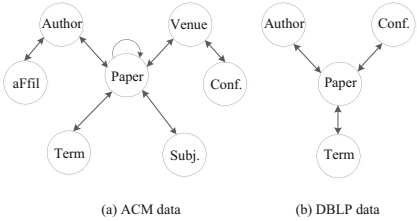


Fig. 1. Bibliographic network schema

networks have been extensively studied in the past decades, such as PageRank [3] and SimRank [4], the similarity measure in heterogeneous network is just beginning now and several measures have been proposed including PathSim [5], PCRW [6] and HeteSim [7]. All the three methods are based on **Meta-Path** whose definition can be found in the related work [7]. Specially, HeteSim, proposed by Shi et al., has the ability to measure relatedness of objects with the same or different types in a uniform framework. HeteSim has some good properties (e.g., self-maximum and symmetric), and has shown its potential in several data mining tasks. However, we can also find that it has several disadvantages. (1) HeteSim has relatively high computational complexity, in particular, the adoption of path decomposition approach while measuring the relevance on odd-length path further increases complexity of calculation. (2) Besides, HeteSim cannot be extended to large-scale network with massive data, since its calculation process is based on memory computing. Therefore, it is desired to design a new similarity measure, which not only contains some good properties of HeteSim but also overcomes the disadvantages on computation.

In this paper, we propose a new relevance measure method - **AvgSim**, which is a symmetric and uniform measure to evaluate the relevance of same or different-typed objects. Since AvgSim can also measure the relevance of different-typed objects, we use the relevance measure instead of similarity measure in the following section. AvgSim value of two objects is the average of reachable probability under the given path and the reverse path. It guarantees that AvgSim can measure relevance of same or different-typed objects and it has symmetric property. In addition, we take parallelization of this new algorithm on MapReduce in order to eliminate restriction of memory size and deal with massive data more efficiently in practical applications. Experiments on real dataset show that AvgSim can achieve comparative performances with high efficiency and effectiveness, compared with other methods including HeteSim, PathSim and PCRW. Moreover, experiments on large-scale dataset also validate the effectiveness of parallelized AvgSim.

The rest of this paper is organized as follows: Section 2 describes AvgSim in detail. And the method of parallelization of AvgSim is explained in Section 3. Section 4 analyzes performance experiment results of AvgSim to validate its effectiveness and efficiency. And some matrix parallelization experiments are also in this section. Finally we conclude this paper in Section 5.

2 AvgSim: A Novel Relevance Measure

In this section, we will introduce you a new meta-path based relevance measure which is called **AvgSim** and the definition of it is as follows.

Definition 1 ***AvgSim:*** *Given a meta-path P which is defined on the composite relation $R = R_1 \circ R_2 \circ \ldots \circ R_l$, AvgSim between two objects s and t (s is the source object and t is the target object) is:*

$$AvgSim(s,t|P) = \frac{1}{2}[RW(s,t|P) + RW(t,s|P^{-1})] \tag{1}$$

$$RW(s,t|R_1 \circ R_2 \circ \ldots \circ R_l) = \frac{1}{|O(s|R_1)|} \sum_{i=1}^{|O(s|R_1)|} RW(O_i(s|R_1),t|R_2 \circ \ldots \circ R_l) \tag{2}$$

Equation (1) shows the relevance of source object and target object based on meta-path P is the arithmetic mean value of random walk result from s to t along P and reversed random walk result from t to s along P^{-1}. Equation (2) shows the decomposed step of AvgSim, namely the measure of random walk. The measure takes a random walk step by step from starting point s to end point t along path P using iterative method, where $|O(s|R_1)|$ is the out-neighbors of s based on relation R_1. If there is no out-neighbors of s on R_1, then the relevance value of s and t is 0 because s cannot reach t. We need to calculate random walk probabilities for each out-neighbor of s to t iteratively, and then sum them up. Finally the summation should be normalized by the number of out-neighbors to get average relatedness. The stop sign of iteration is that s meets t at t node along P. In contrast to simple random work method, AvgSim shows its comprehensiveness and the effectiveness reflected in later experiments verifies its advantages.

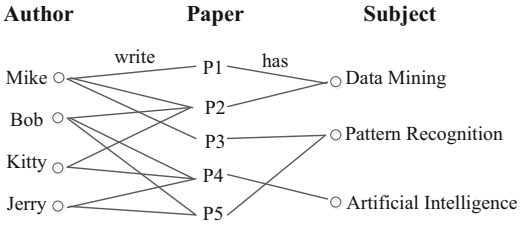


Fig. 2. Heterogeneous relation network example

We take the simple network showed in Fig.2 as an example to calculate the relevance between $Mike$ and the subject $DataMining$ (DM for short) based on path APS ("Author-Paper-Subject ").

$$AvgSim(Mike,DM|APS) = \frac{1}{2}[RW(Mike,DM|APS) + RW(DM,Mike|SPA)] \tag{3}$$

$$RW(Mike, DM|APS) = \frac{1}{|O(Mike|AP)|} \sum_{i=1}^{|O(Mike|AP)|} RW(O_i(Mike|AP), DM|PS) \tag{4}$$

We notice from Fig.2 that $O(Mike|AP) = \{P_1, P_2, P_3\}$, thus we need to calculate relatedness between each out-neighbor of $Mike$ and DM, like $RW(P_1, DM|PS)$.

$$RW(P_1, DM|PS) = \frac{1}{|O(P_1|PS)|} \sum_{i=1}^{|O(P_1|PS)|} RW(O_i(P_1|PS), DM) \tag{5}$$

Since that $O(P_1|PS) = \{DM\}$, out-neighbors of P_1 based on relation PS will meet with DM, thus $RW(P_1, DM|PS) = 1$. Finally, we can easily calculate the relatedness value of random walk from $Mike$ to DM along path APS is 2/3. Likewise, relatedness value of reverse random walk along path SPA is 2/3. Thus the relevance value (i.e. AvgSim) between author $Mike$ and subject $DataMining$ is 0.67 (2/3).

The example above shows the operation process of AvgSim measuring relevance of two arbitrary objects along a meta-path. Next we will study on how to calculate AvgSim generally **using matrices**.

Given a simple directed meta-path $A \xrightarrow{R} B$, where object A and B are linked though relation R. The relationship between A and B can be expressed by adjacent matrix, denoted as M_{AB}. Two normalized matrix R_{AB} and C_{AB} are generated by normalizing M_{AB} according to row vector and column vector respectively. R_{AB} and C_{AB} are **transition probability matrix** which represent $A \xrightarrow{R} B$ and $B \xrightarrow{R^{-1}} A$ respectively. According to properties of matrix, we can derive relations $R_{AB} = C'_{BA}$ and $C_{AB} = R'_{BA}$, where R'_{AB} is the transpose of R_{AB}.

If we extend the simple meta-path to $P = A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \ldots \xrightarrow{R_l} A_{l+1}$ where R is a composite relation $R = R_1 \circ R_2 \circ \ldots \circ R_l$, then the relationship between A_1 and A_{l+1} is expressed as **reachable probability matrix** which is obtained by computation on the basis of transition probability matrix. The reachable probability matrix of P is defined as $RW_P = R_{A_1A_2}R_{A_2A_3}\ldots R_{A_lA_{l+1}}$, where RW suggests RW_P is the random walk relatedness matrix from object A_1 to A_{l+1} along path P.

Then we can rewrite AvgSim using reachable probability matrix according to equation (1) and (2) as follows.

$$\begin{aligned} &AvgSim(A_1, A_{l+1}|P) \\ &= \frac{1}{2}[RW(A_1, A_{l+1}|P) + RW(A_{l+1}, A_1|P^{-1})] = \frac{1}{2}[RW_P + RW'_{P^{-1}}] \end{aligned} \tag{6}$$

Applied relation $C_{AB} = R'_{BA}$, equation (8) is derived below. We notice that the calculation of AvgSim is unified as two chain matrix multiplication of transition probability matrices. The only difference between two chains is the normalization form of original adjacent matrix.

$$AvgSim(A_1, A_{l+1}|P) = \frac{1}{2}[R_{A_1A_2}R_{A_2A_3}\dots R_{A_lA_{l+1}} + (R_{A_{l+1}A_l}R_{A_lA_{l-1}}\dots R_{A_2A_1})^{'}]$$
$$= \frac{1}{2}[R_{A_1A_2}R_{A_2A_3}\dots R_{A_lA_{l+1}} + C_{A_1A_2}C_{A_2A_3}\dots C_{A_lA_{l+1}}] \quad (7)$$

AvgSim can measure relevance of any heterogeneous or homogeneous objects based on symmetrical path ($e.g. APCPA$) or asymmetrical path ($e.g. APS$). Besides, the method has symmetric property, which can be verified easily from the definition equation of AvgSim and the symmetric property has a positive effect on clustering. However, the calculation of AvgSim mainly the chain matrix multiplication is time-consuming and restricted of memory size. In order to apply our algorithm in real large-scale heterogeneous information network, we have to consider how to improve the efficiency of AvgSim.

3 Parallelization of AvgSim

Parallelism is an effective method for processing of massive data and improving algorithm's efficiency. According to the features and application scenarios of AvgSim, we will realize it using parallelization method and the specific steps are as follows.

1. Since the core calculation of AvgSim is the chain matrix multiplication, we firstly change the order of matrix multiplication operations applying Dynamic Programming strategy.

2. After step 1, we turn to focus on single large-scale matrix multiplication and it can be parallelized on Hadoop distributed system using MapReduce programming model.

As we know, different orders of operations in chain matrix multiplication leads to different time of computation. There exists an optimal order of chain matrix multiplication using Dynamic Programming, which consumes the shortest computation time. Thus, we can apply Dynamic Programming to improve the efficiency of parallelized AvgSim. And the parallelization of AvgSim is mainly the parallelization of matrix multiplication after Dynamic Programming process. Here we use "block matrix multiplication "method on MapReduce to transform multiplication of two large matrices into several multiplications of smaller matrices. This method is flexible with selecting dimensions of block matrix according to the configuration of Hadoop cluster and avoids exceeding the memory size.

Applying "block matrix multiplication "iteratively to the chain matrix multiplication which is re-ordered by Dynamic Programming, we can get one of the two reachable probability matrices of AvgSim (e.g., RW_P, which is measured in the given meta-path P), and the other probability matrix ($RW^{'}_{P^{-1}}$) can be obtained in exactly the same procedure. Finally, the relevance matrix is derived by taking arithmetic mean of these two reachable probability matrices.

4 Experiments

4.1 Data Sets

Two data sets, **DBLP dataset** and **Matrix dataset**, are used in experiments and the previous network schema is shown in Fig. 1(b). In detail, the DBLP dataset contains 14K papers, 14K authors, 20 conferences and 8.9K terms. And we label 20 conferences, 100 papers, and 4057 authors in the dataset with four research areas including database, data mining, information retrieval and artificial intelligence for experiments use. And the Matrix dataset *(40 matrices in total)* contains several artificially generated large-scale sparse square matrices, whose dimensions are 1000×1000, 5000×5000, 10000×10000, 20000×20000, 40000×40000, 80000×80000, 100000×100000 and 150000×150000 respectively. And the sparsity of each matrix includes 0.0001, 0.0003, 0.0005, 0.0007 and 0.001.

4.2 Performance of AvgSim

Performance on Query Task and Clustering Task. In the query task, we compare the performance of AvgSim with both HeteSim and PCRW though measuring the relevance of heterogeneous objects on DBLP dataset. Based on labels of the dataset, we calculate the AUC (Area Under ROC Curve) score to evaluate the performance of the results which are the related authors ranked by relevance scores for each conference on meta-path CPA. We evaluated 9 out of 20 marked conferences, whose AUC values are shown in Table 1. We notice that AvgSim gets the highest value on 8 conferences, which means AvgSim performs better than other two methods in the query task.

Table 1. AUC values for relevance search of conferences and authors based on CPA path on DBLP dataset

	KDD	ICDM	SDM	SIGMOD	VLDB	ICDE	AAAI	IJCAI	SIGIR
HeteSim	0.8111	0.6752	**0.6132**	0.7662	0.8262	0.7322	0.8110	0.8754	0.9504
PCRW	0.8030	0.6731	0.6068	0.7588	0.8200	0.7263	0.8067	0.8712	0.9390
AvgSim	**0.8117**	**0.6753**	0.6072	**0.7668**	**0.8274**	**0.7286**	**0.8114**	**0.8764**	**0.9525**

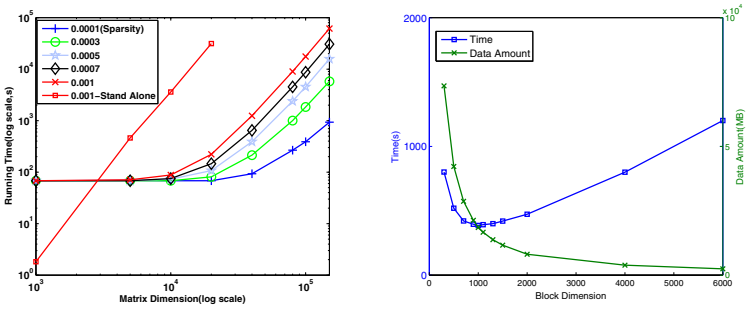
Table 2. Clustering accuracy results for path-based relevance measures on DBLP dataset

	Venue NMI	Author NMI	Paper NMI
PathSim	0.8162	0.6725	0.3833
HeteSim	0.7683	0.7288	0.4989
AvgSim	**0.8977**	**0.7556**	**0.5101**

In the clustering task, we compare the performance of AvgSim with both HeteSim and PathSim though measuring the relevance of homogeneous objects on DBLP dataset. We firstly apply three algorithms respectively to derive

the relevance matrices on three meta-paths including $CPAPC$, $APCPA$ and $PAPCPAP$. Based on the result matrices and applied Normalized Cut, we perform clustering task and then evaluate the performances on conferences, authors, and papers using NMI criterion (Normalized Mutual Information). The clustering accuracy result is shown in Table 2 and AvgSim gets the highest NMI value in all the three tasks. The results of query task and clustering task suggest that AvgSim performs well in effectiveness.

Performance of Parallelized Matrix Multiplication. All parallelized matrix multiplication experiments are conducted on $Matrix$ dataset in a cluster composed of 7 machines with 4-cores E3-1220 V2 CPUs of 3.10GHz and 32 GB RAM running on RedHat 4 operating system. The experiments will measure several factors affecting block matrix multiplication, including matrix dimensions, matrix sparsity and partition strategy (i.e. dimensions of blocks).



(a) Matrix Dimension and Sparsity (b) Partition Strategy

Fig. 3. Factors affecting parallelized block matrix multiplication

Fig.3(a) shows the relationship among matrix dimensions, matrix sparsity and running time of parallelized block matrix multiplication together with the comparison between stand-alone and parallelized matrix multiplication. We notice that the larger dimensions or sparsity of matrix are, the more time in matrix multiplication is required. And the stand-alone algorithm costs shorter time for quite small matrix dimension because parallelized algorithm spends lots of time in starting task nodes of Hadoop cluster and resources of cluster are not fully utilized for small amount of calculations. However, efficiency of parallelized algorithm is much better as matrix dimension increasing. Besides, stand-alone algorithm is restricted of memory size for there are no results derived in the last three large-scale matrix multiplications.

Fig.3(b) shows the relationship among running time, intermediate data amount and partition strategy of block matrix multiplication. There are 11 kinds of partition strategies with square block matrix dimensions of 300×300, 500×500, 700×700, 900×900, 1000×1000, 1100×1100, 1300×1300, 1500×1500, 2000×2000, 4000×4000 and 6000×6000 respectively applying in the square matrix with dimension of 100000×100000 and a sparsity of 0.0001 in the experiment. We notice that intermediate data amount of matrix multiplication decrease

gradually with the increase of block dimension. In contrast, running time reaches its minimum value at 5-th data point shown in figure. Smaller intermediate data amount results in less disk IO operations and data amount transmitted by shuffle, which also means shorter time and better performance to a certain extent as front several data points reflected. However, excessive large block dimension will reduce the concurrent granularity and increase the amount of calculations for single node, which conversely results in longer time of computation as several data points behind reflected.

In conclusion, appropriate partition strategy and sufficient sizes of cluster greatly affect the efficiency in parallelized block matrix multiplications. Applying parallelization method, AvgSim gains the ability to measure relevance in larger-scale networks with massive data efficiently.

5 Conclusions

In this paper, we introduced a novel algorithm with symmetrical features named AvgSim for measuring relevance of arbitrary objects in heterogeneous information network. In addition, using Dynamic Programming and "block matrix multiplication "methods, parallelized AvgSim is able to be applied to actual large-scale networks. Experiments given in the paper verified the effectiveness and efficiency of AvgSim while measuring the relevance of heterogeneous or homogeneous objects based on meta-paths.

Acknowledgment. This work is supported by the National Key Basic Research and Department(973) Program of China (No.2013CB329603), the National Science Foundation of China (Nos.61375058, and 71231002), the Ministry of Education of China and China Mobile Research Fund (MCM20123021) and the Special Co-construction Project of Beijing Municipal Commission of Education.

References

1. Sun, Y., Han, J., Zhao, P., Yin, Z., Cheng, H., Wu, T.: RankClus: integrating clustering with ranking for heterogeneous information network analysis. In: EDBT, pp. 565–576 (2009)
2. Kong, X., Yu, P.S., Ding, Y., Wild, D.J.: Meta path-based collective classification in heterogeneous information networks. In: CIKM, pp. 1567–1571 (2012)
3. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: bringing order to the web. Stanford University Database Group. Technical report (1998)
4. Jeh, G., Widom, J.: SimRank: a measure of structural-context similarity. In: KDD, pp. 538–543 (2002)
5. Sun, Y., Han, J., Yan, X., Yu, P., Wu, T.: Pathsim: meta path-based top-k similarity search in heterogeneous information networks. In: VLDB, pp. 992–1003 (2011)
6. Lao, N., Cohen, W.: Relational retrieval using a combination of path-constrained random walks. Machine Learning 81(1), 53–67 (2010)
7. Shi, C., Kong, X., Huang, Y., Yu, P.S., Wu, B.: HeteSim: A General Framework for Relevance Measure in Heterogeneous Networks. In: CoRR, pp.abs/1309.7393 (2013)