ISSN (online) : 2277-5390
ISSN (Print) : 2277-5382

A Survey on Different Similarity Join To Improve Clustering, Classification and Similarity Search

C.P. Rushida*[1], V R Nagarajan[2]

[1]*Student, M.Phil.in Computer Science, Sri Narayana Guru College, Coimbatore-641105, Tamil Nadu, India*
[2]*Assistant Professor, M. Phil. in Computer Science, Sri Narayana Guru College, Coimbatore-641105, Tamil Nadu, India*
rushidacp@yahoo.in[1], vrnag74@gmail.com[2]

Abstract

Objectives: To analysis various similarity join techniques to improve the data mining process.
Findings: Similarity join is an evaluation of similarity between any two objects. Many applications such as data cleaning, data integration, near duplicate detection and all data mining process can extensively benefit from the similarity join measure. Thus the similarity join can be performed between objects or strings or nodes etc. It finds all pairs of objects whose similarity is not smaller than the similarity threshold. There are different techniques and approaches are used to find the similarity join between objects in homogeneous information network. This paper provides detailed information about the different similarity join techniques.
Results: In this paper various similarity join techniques are compared through parameters to prove path based similarity join is better than other techniques.
Application/Improvements: The findings of this work prove that the path based similarity join provides better result than other approaches.

Keywords: similarity join, data cleaning, data integration, near duplicate detection.

1. Introduction

Rapid growth of data on the internet leads to define the similarity between objects which is an essential operation for many applications. Recently, it received a significant attention in both academic and industrial community. The similarity can be defined even between the strings which find the similar strings between pair two set of strings which finds the near duplicated queries. Additionally, inter related information are available in the internet it is difficult task to define the similarity between the objects. There are two types of similarity measures are available to define the similarity join. One is content based similarity measure it treats each objects as bag of items and another one is link based similarity measure which consider object to object relationship it is expressed in terms of links. The similarity join is an essential operation in various domain and applications. There were various techniques developed for this purpose.

Various techniques effectively find the similarity join for data integration, clustering [1], and filtering. In link based similarity join understand the relationship between all pair of nodes with the help of Personalized PageRank and SimRank measure. Pass Join method finds the similar string by partitioning the string and selects the suitable substring for pruning techniques that minimize the quantity of selected substrings and it analysis the candidate pairs. Whereas in near duplicate detection, near duplicate records are found by filtering techniques which exploit the ordering information and time sequences records. The similarity join can be determined by using local sensitive hashing and MapReduce techniques. Path based similarity join is build up to give top k similarity pairs in heterogeneous information network and it derives various similarity semantics.

In [2] investigated the problems involved in probabilistic set similarity join in probabilistic set database and proposed effective pruning techniques for the problems. The pruning techniques effectively filter out the false pairs of probabilistic set it automatically reduces probabilistic set similarity join search space. In this M-tree index built upon probabilistic set data to make easy the probabilistic set similarity join processing. Further designed an outline for concluding probabilistic set data, which can be flawlessly integrated into the developed index and facilitate the pruning with the probabilistic threshold.

In [3] proposed LSH-SS algorithm that is sampling based algorithm for similarity join. It used Locality Sensitive Hashing (LSH) index to facilitate effective sampling even at high thresholds it can be applied across various domains. The main goal of this algorithm is that though sampling a pair providing a high threshold is very difficult, it is comparatively easy to sample the pair with the help of LSH scheme because it creates group with similar objects and it provides good estimates throughout the similar threshold range with a sample size of and pairs of vectors with probabilistic guarantees.

In [4] proposed non iterative computation technique which is an optimization approach for SimRank measure. In the proposed technique vectorization operators and Kronecker product is used to rewrite the SimRank equation in a non iterative form. It creates similarity scores by computing individual nodes in a network in a linear time without computing the similarity matrix. When the network changes eventually, this technique provides any-time query answer by updating SimRank scores incrementally and two more algorithms are proposed called C_Track and S_Track to analysis centrality tracking and node similarity respectively to improve the performance of proposed methods. It can be used for both static and dynamic information network.

In [5] proposed a framework called trie-join to improve the performance of similarity join with edit distance constraint. This framework used trie structure to index the string and used an effective pruning technique to reduce the candidate pair which utilizes the trie structure to find the similar string. This framework achieves high performance by dynamic update of datasets.

In [6] presented link based similarity join (LS-join) for consideration the relationship between all pair of nodes facilitates application. LS-join obtains the input as pair of nodes and it will produce all pair of nodes. Then IDJ algorithm is proposed to estimate the LS-join, which can be utilized on a class of link-based measures and improves the performance of the IDJ for the Personalized PageRank and the SimRank, which are frequently used similarity measures. These algorithms, which can be processed on directed and weighted graphs, performed better than basic solutions on large graphs.

In [7] introduced new approach for similarity measure between two objects called Meta path-based similarity. There are different measures are available for similarity measure are SimRank on the extracted sub-network, pair wise random walk used in SimRank, or directly apply, random walk in P-PageRank and P-PageRank. These measures determine the similar objects are based upon highly concentrated objects. But in the proposed measure namely PathSim measure able to confine the delicate semantics of similarity among examine objects in a network and it can determine objects share similar visibility in the network along with the strongly connected objects given the Meta path.

In [8] proposed a new filtering technique called positional filtering to decrease the candidate sizes in similarity join. It makes use of the organizing of tokens in a record and provides to upper bound estimates of similarity scores. It shows that it is equivalent to the existing prefix filtering method and can works on tokens both in the suffixes and the prefixes. On the other hand it successfully eases the problem of quadratic growth of candidate pairs when the data increases in size.

In [9] proposed fuzzy token matching based similarity which a new similarity metric. This metric comprises of character similarity and token similarity. In token based similarity defines the similarity based on exact match between tokens it also integrates character-based similarity of mismatched token pairs into the fuzzy-token similarity. There are several problems occurred due to fuzzy-token similarity and proposed a new framework called signature-based framework which is used to address such problems where developed a new signature scheme for tokens and present effective punning techniques to improve the performance.

In [10] proposed a new method called trie join for string similarity join. It overcome problems like to find the similarity join in short string, large indexes and dynamic update of dataset. Trie join produces the result with small number of indexes through the trie structure which finds similar strings based on subtrie pruning. It occupies much smaller space to hold the indexes of string and the performance of subtrie pruning is improved by dual substring pruning technique from this technique the large number of strings was prune. Hence it avoids unnecessary computation through count filtering and length filtering techniques.

In [11] proposed Versatile Scalable MapReduce all-pair similarity Join (V-SMART JOIN) which provides best solution for timely problems. In these method two stage algorithms is used for internet traffic. In the first stage of algorithm computes and join partial results whereas in the second stage calculates the similarity accurately for all candidate pairs. This is used for classify IP addresses as load balancing proxies. It effectively process on sets, vectors and multi sets with a huge variety of similarity measures and it handles skewed data distributions.

In [12] presented a new method for similarity measure of all pairs of objects called Bayes Locality-sensitive hashing (BayesLSH). BayesLSH-Lite, which calculates similarities exactly, is also presented. BayesLSH-Lite algorithms improve speedups over baseline approaches by remove away huge number of false positive candidate pairs. Additionally it offered probabilistic guarantees over the output in terms of accuracy and recall. Further the BayesLSH output quality can be easily adjusted and this method there is no need for any manual settings for number of hashes used in hash function to estimate their similarity. The advantages of BayesLSH are repaying the costs of hashing, it demonstrated for Jaccard and Cosine similarity measure, there is no assumption for candidate generation algorithm, and it performs well for binary and general real valued vectors and parameter tuning is very comfortable and spontaneous.

In [13]proposed a new solution called HeteSim for the problems involved in the similarity measure between objects of same type and different type in heterogeneous network which is a novel similarity measure. It measure the similarity between the objects with the help of path constrained measure, uniform measure and semi metric measure. The measure which measures the relatedness of object pairs are characterized based on the search path that join two objects through a series of node types called path constrained measure. The measure which measures the relatedness of objects in a uniform framework includes the any types of objects called uniform measure and semi metric measure has been applied for many data mining process.

In [14] proposed a new approach for the purpose of clustering objects in heterogeneous information network that merge the Meta path selection and user-guided clustering. In this proposed approach initially user initialize seeds for each cluster it provides guidance for each cluster in the network. The cluster is fully depends or based on initialized seeds that learned the weights for every Meta path to cluster the objects. Here some issues will be raised due to initialized seeds doesn't provide efficient clustering and it is overcome by proposed probabilistic approach additionally an effective and efficient iterative algorithm called PathSelCluswas developed where the clustering quality and the meta-path weights are equally improving each other.

In [15] proposed new method called partition-based method for similarity join of both short string and long string. This method divides the input string into number of subset of strings (segments) then develops inverted indices for segmented strings. Based on the inverted indices the candidate pair of each string is found by selecting some of the substrings from the input string and construct efficient methods to select the substrings which decrease the number of selected substrings. The candidate pairs are verified based on the proposed extension based method and its performance can be improved by developing early termination techniques and pruning techniques.

In [16] proposed a new adaptive framework for similarity join in data cleansing processing. The different objects should have the different prefix length prefix filtering is used for this purpose. Hence in this method cost model is proposed which choose a suitable prefix for each object additionally developed effective indexes for each object that helps to select the proper prefix. It supports all similarity functions and it showed high performance in similarity search and similarity join.

In [17] proposed an approach with local sensitive hashing and MapReduce for time sequences similarity join. Initially Discrete Fourier Transform is used to convert each time sequence into frequency domain. Then find the candidate similar time sequence pair with the help of Locality Sensitive Hash which reduces the computation of sequence pairs through only consider similarity pairs and in this step duplicated pairs are also removed. Finally MapReduce framework is used to improve the performance of similarity join in massive amount of time sequences.

In [18] proposed a new measure called HeteSim for relevance search problem in heterogeneous networks and proposed computation strategies for HeteSim measure. This measure works more effectively in data mining process. This measure has the characteristics are uniform measure, path constrained measure and semi-metric measure. Uniform measure is used to measure the relatedness of objects with the same or different types in a uniform framework then the second measure path constrained measure based on the search path between the object pairs which utilized sequences node types to connect the object pairs and semi-metric measure is used for data mining process. Thus the symmetric measures are more used in learning tasks and in many applications.

In [19] proposed h-go score index to determine SimRank score. In this method focused on SimRank-based join query problem over large graphs and provided a non-iterative computation model to calculate the SimRank. It reduces the search space by determined shortest-path distance based upper bound for SimRank scores to reduce glooming vertex pairs and then used h go score index to calculate SimRank score.

In [20] proposed a solution for capably determining important meta-paths in large knowledge bases and developed the Forward Stagewise Path Generation algorithm (or FSPG), which originates meta-paths that most excellent predict the similarity between the given node pairs through greedy algorithm. FSPG is that it supports existing meta-path-based similarity measures Path Count and Path Constraint Random Walk. These measures are used for determine the Meta path between nodes can be used for decision making, link prediction and product recommendation.

In [21] proposed path based similarity join for data mining purpose in heterogeneous information network. This path based similarity join used nearby buckets to create candidate object pairs. It reduces the similarity computation between buckets by discarding pairs that are located far away buckets. This method is enhanced by capture a similar pair hashed into different buckets, then expands the Locality Sensitive Hashing (LSH) table with an additional bucket array and utilizes this information to reduce the buckets that do not need to be compared. This method is called as bucket pruning based Locality Sensitive Hashing gives more effective solution for top k similarity join.

2. Comparison of similarity join techniques

Ref. No	Title	Parameters Used
1	Set Similarity Join on Probabilistic Data	Speed up ratio NLJ (3% Probabilistic elements)=0.0053 PS^2J(3% Probabilistic elements)=0.0046
2	Similarity Join Size Estimation using Locality Sensitive Hashing	Relative error % LSH-SS (0.5 similarity threshold) =23 LSH-SS (D)(0.5 similarity threshold) =27 RS (cross) (0.5 similarity threshold) = 42
3	Fast Computation of SimRank for Static and Dynamic Information Networks	Accuracy NDCG@5 = 94% NDCG@10 = 92%
4	TrieJoin: Efficient Trie based String Similarity Joins with Edit Distance Constraints	Index size (Mb) Trie-PathStack (AOL) =29 Part-Enum (AOL) = 120 All-Pairs-Ed (AOL) = 305

5	On Link-based Similarity Join	Running time (200 nodes)
		IDJ-UB1= 200 secs
		IDJ-UB2= 80 secs

6	Pathselclus: Integrating Meta-Path Selection with User-Guided Object Clustering in Heterogeneous Information Networks	Accuracy for PathSelClus = 0.9950 Accuracy for LP = 0.6500 normalized mutual information for PathSelClus = 0.9906 normalized mutual information for LP = 0.6181
7	Efficient Similarity Joins for Near Duplicate Detection	Precision at t=0.95=0.38 Recall at t=0.95=0.11 Precision at t=0.90=0.48 Recall at t=0.90=0.06

8	Fast-Join: An Efficient Method for Fuzzy Token Matching based String Similarity Join	Precision
		fuzzy-jaccard =95%
		fuzzy-dice =71%

9	Trie-join: a trie-based method for efficient string similarity joins	Index sizes(MB) Trie-PathStack (AOL Query Log) = 29 Bi-Trie-PathStack(AOL Query Log) = 80 Part-Enum (AOL Query Log) = 120 All-Pairs-Ed PathStack(AOL Query Log) = 305
10	V-SMART-Join: A Scalable MapReduce Framework for All-Pair Similarity Joins of Multisets and Vectors	Run time (secs) VCL algorithm=5000 secs (at 0.8 similarity threshold) V-SMART-Join= 1200 secs(at 0.8 similarity threshold)
11	Bayesian Locality Sensitive Hashing for Fast Similarity Search	Recall Bayes-LSH (t=0.5) = 97.97 (RCV1) Bayes-LSH (t=0.6) = 98.52 (WikiWords100K) Bayes-LSH-Lite(t=0.5) = 98.73 (RCV1) Bayes-LSH-Lite (t=0.6) = 98.88 (WikiWords100K)
12	Relevance Search in Heterogeneous Networks	Clustering accuracy HeteSim (Author NMI) = 0.7288 HeteSim (Paper NMI) = 0.4989 PathSim (Author NMI) = 0.8162 PathSim (Paper NMI) = 0.3833

13	PathSim: Meta PathBasedTopK Similarity Search in Heterogeneous Information Networks	Accuracy
		P-PageRank = 0.5552
		SimRank = 0.6289
		RW = 0.7061
		PathSim = 0.7446

14	Pass-Join: A Partition based Method for Similarity Joins	Index size (MB) PASS-JOIN (author) = 1.92 ED-JOIN(author) = 25.34

15	Can We Beat the Prefix Filtering? An Adaptive Framework for Similarity join and search	Running time (secs) Adapt-join(at 0.5 threshold)=2000 Pp join (at 0.5 threshold)=1800 Ppjoin+ (at 0.5 threshold)=1200
16	Efficient Similarity Join for Massive Time Sequences Using Locality Sensitive Hash and MapReduce	Execution time(sec) For 50000 time sequence=906 secs For 100000 time sequence =4450 For 200000 time sequence =7821secs
17	HeteSim: A General Framework for Relevance Measure in Heterogeneous Networks	AUC HeteSim (KDD) = 0.8111 PCRW (KDD) = 0.8030 HeteSim (ICDM) = 0.6731
18	Efficient SimRank-based Similarity Join Over Large Graphs	Index building time (secs) for SRJ ER100K dataset=343.136 Yeast dataset= 80.316 Cora dataset =1617.356
19	Discovering Meta-Paths in Large Heterogeneous Information Networks	True positive rate (Yago-advisorOf) at 0.2 false positive rate FSPG= 0.4 PCRW 2= 0.25 PCRW 4=0.2 PCRW 3 = 1.8

20	Top-k Similarity Join in Heterogeneous Information Networks	Recall for k=1000
		LSH=0.99
		NBLSH=0.99
		BPLSH=1
		Error ratio for k=1000
		LSH=1.00003
		NBLSH=1.00003
		BPLSH=1
		Time(s) for k=1000
		LSH= 6.592
		NBLSH= 8.988
		BPLSH=5.473
		Hash tables for k=1000
		LSH= 5
		NBLSH= 2
		BPLSH=1

3. Conclusion

Various techniques are available to find similarity join between objects. Among them, path based similarity join method returns top k similar pairs of objects in heterogeneous information network and it gives various similarity semantics by pruning and optimization techniques.

4. References

1. R. Nagaraj, V. Thiagarasu, B. Jeevithapriya. Optimization and scalable constrained clustering performances. *Indian Journal of Innovations and Developments.* 2015;4(7),1-7.
2. X. Lian, L. Chen. Set similarity join on probabilistic data.Very Large Database *Endowment*. 2010; 3(1-2), 650-659.
3. H. Lee, R. T. Ng, K. Shim. Similarity join size estimation using locality sensitive hashing. Very Large Database *Endowment Endowment*. 2011; 4(6), 338-349.
4. C. Li, J. Han, G. He, X. Jin, Y. Sun, Y. Yu, T. Wu. Fast computation of simrank for static and dynamic information networks. *International Conference on Extending Database Technology*. 2010; 465-476.
5. J. Wang, J. Feng, G. Li. Trie-join: Efficient trie-based string similarity joins with edit-distance constraints. Very Large Database *Endowment*. 2010; 3(1-2), 1219-1230.
6. L. Sun, C. K. Cheng, X. Li, D. W. L. Cheung, J. Han.On link-based similarity join. Very Large Database *Endowment*. 2011; 4(11), 714-725.

7. Y. Sun, J. Han, X. Yan, P. S. Yu, T. Wu. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks.Very Large Database *Endowment.* 2011; 4(11), 992-1003.
8. C. Xiao, W. Wang, X. Lin, J. X. Yu, G. Wang. Efficient similarity joins for near-duplicate detection. *ACM Transactions on Database Systems (TODS)*. 2011; 36(3), 131-140.
9. J. Wang, G. Li, J. Fe. Fast-join: An efficient method for fuzzy token matching based string similarity join. *International Conference on Data Engineering*. 2011; 458-469.
10. J. Feng, J. Wang, G. Li. Trie-join: a trie-based method for efficient string similarity joins. *The VLDB Journal-The International Journal on Very Large Data Bases*. 2012; 21(4), 437-461.
11. A. Metwally, C. Faloutsos. V-smart-join: A scalable mapreduce framework for all-pair similarity joins of multisets and vectors. Very Large Database *Endowment*. 2012; 5(8), 704-715.
12. V. Satuluri, S. Parthasarathy. Bayesian locality sensitive hashing for fast similarity search. Very Large Database *Endowment*. 2012; 5(5), 430-441.
13. C. Shi, X. Kong, P. S. Yu, S. Xie, B. Wu. Relevance search in heterogeneous networks. In *Proceedings of the 15th International Conference on Extending Database Technology*. 2012; 180-191.
14. Y. Sun, B. Norick, J. Han, X. Yan, P. S. Yu, X .Yu. Pathselclus: Integrating meta-path selection with user-guided object clustering in heterogeneous information networks. *ACM Transactions on Knowledge Discovery from Data (TKDD)*. 2013; 7(3), 1348-1356.
15. G. Li, D. Deng, J. Wang, J Feng. Pass-join: A partition-based method for similarity joins. Very Large Database *Endowment*. 2011; 5(3), 253-264.
16. J. Wang, G. Li, J. Feng. Can we beat the prefix filtering?: an adaptive framework for similarity join and search. *ACM SIGMOD International Conference on Management of Data* 2012; 85-96.
17. D. Chen, L. Zheng, M. Zhou, S. Yu. Efficient Similarity Join for Time Sequences Using Locality Sensitive Hash and Mapreduce. In *Cloud Computing and Big Data (CloudCom-Asia),* 2013;529-533.
18. C. Shi, X. Kong, Y. Huang, S. Y. Philip, B. Wu. Hetesim: A general framework for relevance measure in heterogeneous networks. *IEEE Transactions on Knowledge and Data Engineering*, 2014; 26(10), 2479-2492.
19. W. Zheng, L. Zou, Y. Feng, L. Chen, D. Zhao. Efficient simrank-based similarity join over large graphs. *Very Large Databse Endowment*. 2013; 6(7), 493-504.
20. C. Meng, R. Cheng, S. Maniu, P. Senellart, W. Zhang. Discovering meta-paths in large heterogeneous information networks. *International Conference on World Wide Web*. 2015; 754-764.
21. Y. Xiong, Y. Zhu, S. Y. Philip.Top-k similarity join in heterogeneous information networks. *IEEE Transactions on Knowledge and Data Engineering*. 2015; 27(6), 1710-1723.

The Publication fee is defrayed by Indian Society for Education and Environment (iSee). www.iseeadyar.org

Citation:

C.P. Rushidha, V R Nagarajan. A Survey on Different Similarity Join To Improve Clustering, Classification and Similarity Search. *Indian Journal of Innovations and Developments.* 2016; 5(5), May.