User Guided Entity Similarity Search Using Meta-Path Selection in Heterogeneous Information Networks

Xiao Yu, Yizhou Sun, Brandon Norick, Tiancheng Mao, Jiawei Han
Computer Science Department
University of Illinois, at Urbana-Champaign
201 N Goodwin Ave, Urbana IL, US, 61801
{xiaoyu1, sun22, bnorick, mao5, hanj}@illinois.edu

ABSTRACT

With the emergence of web-based social and information applications, entity similarity search in information networks, aiming to find entities with high similarity to a given query entity, has gained wide attention. However, due to the diverse semantic meanings in heterogeneous information networks, which contain multi-typed entities and relationships, similarity measurement can be ambiguous without context. In this paper, we investigate entity similarity search and the resulting ambiguity problems in heterogeneous information networks. We propose to use a meta-path-based ranking model ensemble to represent semantic meanings for similarity queries, exploit the possibility of using using user-guidance to understand users query. Experiments on real-world datasets show that our framework significantly outperforms competitor methods.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data Mining

General Terms

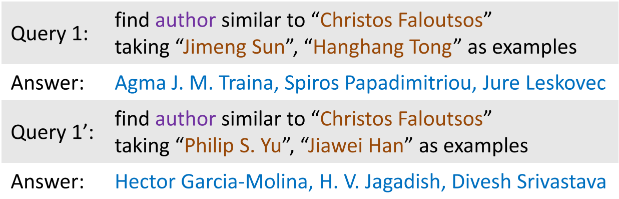
Algorithms, Experimentation

Keywords

Entity Similarity Search, User Guided, Heterogeneous Information Network

1. INTRODUCTION

Due to the rapid emergence of social networks and information networks extracted from various on-line databases, heterogeneous information networks are becoming ubiquitous, which usually contain a large number of multi-typed entities and links (representing entity relationships). We study entity similarity search, which takes entities (nodes) as examples in the query, and returns relevant entities by measuring similarity across the network. Previous studies on query answering in graphs or networks usually follow traditional learning-to-rank procedure and build ranking models by utilizing similarity measurements [2, 6], and these studies usually focus on one specific search task, e.g., friends recommendation [11], or co-authorship prediction [8]. A unified entity similarity search framework, which can understand the similarity semantic differences among queries, and dispatch queries to the proper ranking models automatically, is more desirable than single purposed vertical search systems. To build such a similarity semantic meaning aware search system, we need to first study the similarity semantic ambiguity problem in heterogeneous information networks.



Query 1:	find author similar to "Christos Faloutsos" taking "Jimeng Sun", "Hanghang Tong" as examples
Answer:	Agma J. M. Traina, Spiros Papadimitriou, Jure Leskovec
Query 1':	find author similar to "Christos Faloutsos" taking "Philip S. Yu", "Jiawei Han" as examples
Answer:	Hector Garcia-Molina, H. V. Jagadish, Divesh Srivastava

Figure 1: Different Similarity Semantic Meanings

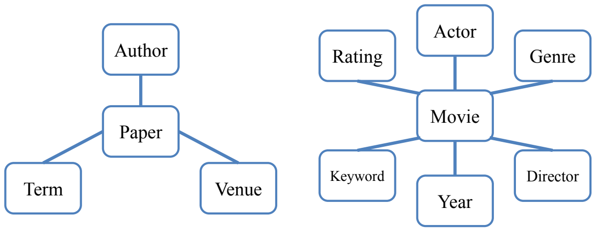
An example of similarity query ambiguity can be found in Figure 1. From the example, one can tell similarity queries which share the same format can possess different semantic meanings. For instance, both Query 1 and Query 1' aim to find authors similar to "Christos Faloutsos", however, if we use the same ranking function to answer both queries, the results might not be satisfactory. In Query 1, the hidden similarity semantic meaning is described by two author entities "Jimeng Sun" and "Hanghang Tong". Judged with human knowledge, both authors were frequent collaborators of Faloutsos. When users issue Query 1, they probably are expecting to find other collaborators of Faloutsos. However, in Query 1' the similarity semantic meaning is entirely different. Users who issue Query 1' might be looking for other highly reputed data mining researchers similar to Faloutsos. We argue that a single ranking model cannot accommodate the range of possible semantic meanings for similarity queries, thus it cannot solve the query ambiguity problem. Although similar problems have been studied in text search engines [5], to the best of our knowledge, solutions to similarity query semantic ambiguity problem have not been proposed in the scope of heterogeneous information networks before.

The major contributions of this paper are summarized below.

- We study the similarity query ambiguity problem in heterogeneous information networks, and propose to use meta-path-based similarity feature space to interpret different similarity semantic meanings.

- We design entity ranking model ensemble and query dispatcher which can choose semantically matched ranking models for a given query, thus produce better results.

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.
Copyright 2012 ACM 978-1-4503-1156-4/12/10 ...$10.00.



(a) DBLP Schema (b) IMDb Schema

Figure 2: DBLP and IMDb Network Schema

2. PROBLEM DEFINITION AND FEATURE SPACE

We use same definitions and notations as in our previous works [9] and [10]. Network schema for the DBLP and IMDb networks are shown in Figure 2. User guided entity similarity search problem is defined in the scope of heterogeneous information network, which is information network with multi-typed entities and links.

DEFINITION 1 (USER GUIDED ENTITY SIMILARITY SEARCH). *Given a heterogeneous information network G, the entity similarity search problem is to find a list of entities similar to a query entity, q. Besides q, users can also provide one or more entities (examples) of the same type as guidance, which usually carries the similarity semantic meaning implicitly.*

One can notice that, a desired search engine which can solve the user guided entity similarity search problem should be able to first understand the hidden semantic meanings of the query and the examples, and then answer the query within the scope of such semantic meanings.

Meta-paths are paths between entities in the network schema of a heterogeneous information network (e.g., $paper \xrightarrow{uses} term \xrightarrow{used} paper$ is a meta-path in the DBLP network). Previous studies [9] [10] [7] suggest that different meta-paths convey different semantic meanings between entities. Other works [6] [2] propose similarity measurements over networks, which calculate distance / similarity between entities quantitatively using different structure heuristics.

The meta-path-based feature space is a combination of meta-paths with meta-path-based similarity measures, i.e., $\mathbf{F} = \mathbf{P} \times \mathbf{M}$, where $\mathbf{P}$ is the set of possible meta-paths and $\mathbf{M}$ is the set of possible meta-path-based measurements.

3. SIMILARITY SEMANTIC MEANING AWARE RANKING MODELS

With the meta-path-based similarity feature space, we now have methods to represent different entity similarity semantic meanings, and with these features we can now measure entity similarity from different semantic perspectives quantitatively.

3.1 Ranking Model Definition

In order to learn a high quality ranking model for each similarity semantic meaning, instead of using the entire meta-path-based similarity feature space, we first apply feature selection process to generate a feature subspace which contains only features relevant to the target semantic meaning. We then build a linear ranking model using the selected meta-path-based features for each semantic meaning, and learn the parameters using training datasets. We define the meta-path-based linear combination ranking model for semantic meaning i as follows.

$$S_i(F^*) = \sum_{f \in \mathbf{F}^*} f \cdot \theta_f \tag{3.1}$$

where $\mathbf{F}^*$ is a selected subset of meta-path-based feature space, and θ_f is the coefficient associated with each feature f in $\mathbf{F}^*$. With this standard linear ranking model, similarity semantics hidden in the training dataset can be captured and represented by a set of meta-path-based features. Coefficients associated with features indicate the importance and expressiveness of each feature in terms of describing the target similarity semantics.

3.2 Training Dataset and Feature Selection

Before feature selection and ranking model training, we first need to prepare a training dataset for each similarity semantic meaning. Training data instances are a set of similar (positive) and dissimilar (negative) entity pairs in the heterogeneous information network according to the given semantic meaning.

Training datasets in this paper are prepared following a *per-query* format, i.e., for each entity, we collect a number of labeled entities (both positive and negative) under the similarity semantic meaning. The training dataset can be formalized as follows.

$$\mathcal{D} = (q_i, q_i^+, q_i^-), i = 1, \ldots, N \tag{3.2}$$

where q_i is the query entity, q_i^+ is a set of the positive examples representing the correct answers (relevant entities) to query q_i, while q_i^- is a set of the negative examples representing the incorrect answers (irrelevant entities) to q_i under the similarity semantic meaning.

According to [4], information gain or entropy based feature selection metrics are not suitable for picking up relevant features for ranking problems. Based on this observation, we propose a feature selection method for the our ranking model, which is similar to the widely used Kendall tau rank correlation coefficient [1].

Given a training dataset and a comprehensive meta-path-based feature space, for each query q_i in the dataset, we first enumerate all positive and negative pairs by calculating $q_i^+ \times q_i^-$. High quality features should be able to rank positive examples higher than negative ones. In order to rank two instances correctly, a feature f should have $f(q_i, q_i^{+(m)}) > f(q_i, q_i^{-(n)})$ for a given query q_i and a positive negative pair $(q_i^{+(m)}, q_i^{-(n)})$. Correct ranked positive and negative pairs are denoted as *concordant pairs* for feature f, and incorrect ranking pairs are denoted as *discordant pairs* for feature f. Notice that, pairs which have $f(q_i, q_i^{+(m)}) = f(q_i, q_i^{-(n)})$ are neither concordant nor discordant.

Rather than using the original Kendall tau rank coefficient to measure the correlation between the feature and the training dataset, we propose to use concordant and discordant ratio to rank features, due to different scoring abilities for meta-path-based features, i.e., the derived similarity matrix could be very sparse for some feature while very dense for others. Features with high scoring ability do not necessarily have a high ranking ability since the ranking results provided by these features can be incorrect.

Kendall tau rank coefficient favors features with high scoring ability while concordant and discordant ratio does not have such bias. By summing concordant and discordant ratios together, we can define the correlation between a meta-path-based feature f and training dataset D as follows.

$$rank\_corr(f, D) = \sum_{q \in \mathcal{D}} \frac{concordant(q, f) + 1}{discordant(q, f) + 1} \tag{3.3}$$

After calculating the ranking correlation of each feature and the training dataset, rather than choosing top-K most relevant features, we employ an entropy based histogram thresholding method to finally decide which features to use. More details can be found in [3].

3.3 Ranking Model Learning

Ranking model learning using the meta-path-based feature space follows a similar procedure to the traditional learning-to-rank task.

We here use a similar objective function to that proposed in [7]. In order to capture the similarity hidden in the training dataset, the goal is to learn the weights associated with each feature that are able to correctly separate as many positive and negative example pairs as possible, for all the queries in the training dataset.

The objective function is defined as follows:

$$O(\theta) = \sum_{i=1,\dots,N} o_i(\theta) \tag{3.4}$$

where $o_i(\theta)$ is per-query objective function, measuring the performance of ranking model on q_i. Per-query objective function $o_i(\theta)$ is defined as log-likelihood function for all positive and negative instances related to q_i as follows:

$$o_i(\theta) = \sum_{m \in q_i^+} \frac{ln(p_i(m))}{|q_i^+|} + \sum_{m \in q_i^-} \frac{ln(1 - p_i(m))}{|q_i^-|} \tag{3.5}$$

where $p_i(m) = \sigma(\theta F^*)$ and $\sigma(x) = \frac{e^x}{1+e^x}$ is the sigmoid function. Potential bias due to an unequal number of positive and negative instances is resolved by normalizing positive instances and negative instances so that their impacts on the objective function are the same. Equation 3.4, the objective function, is a sum over all queries in the training dataset. By maximizing this function, a linear combination ranking model with consistently good performance can be learned with a standard optimization method. We use Broyden Fletcher Goldfarb Shanno method in our experiments.

4. SIMILARITY QUERY DISPATCHER

With the method introduced in last section, given sufficient training datasets of different similarity semantic meanings, we can now build a ranking model ensemble. Each ranking model can answer similarity queries under the scope of a specific similarity semantic meaning. Given an entity similarity query, we do not need to answer this query with all ranking models, but we should with only the ones matching the semantic meanings of the similarity query. In this case, we need a module which can understand or infer the hidden semantic meaning of a given query, and refine the ranking model ensemble by picking up the ranking models which matches the similarity semantic meaning of the query, and only utilize such ranking models to finally answer the query.

4.1 Training Data Preparation

The training dataset for the query dispatcher can be generated from ranking model training dataset $\mathcal{D}$ with re-organizing the dataset by result. The intuition is: queries that share the same relevant results should have similar semantic meanings, while queries that lead to different results should have different semantic meanings. For a candidate retrieval result r in $\mathcal{D}$, positively related queries are collected and denoted as S_r^+, where "positively related" means that under a fixed semantic meaning, r would be retrieved and ranked as a relevant entity for a user input query q, i.e., $r \in q^+$. Similarly, we denote S_r^- as queries which are negatively related to entity r. The query dispatcher should be able to distinguish S_r^+ and S_r^- for a given r. Based on the re-organization, positive training examples are queries which could be used to retrieve same entities as relevant search result, while negative training examples are query pairs which have opposite relevance towards one possible result. The query dispatcher training dataset is denoted as $\mathcal{D}^\mathcal{T}$ in this paper (Equation 4.6).

$$\mathcal{D}^\mathcal{T} = r_i^+, r_i^-, i = 1, \dots, N \tag{4.6}$$

where $r^+ = S_r^+ \times S_r^+$ and $r^- = S_r^+ \times S_r^-$.

4.2 Query Dispatcher Model Learning

After re-organizing the training dataset, the query dispatcher model can now be defined and learned on the new training dataset. Similarity semantic meanings in heterogeneous information networks are proposed to capture the purpose and motivation of queries, and these motivations usually have different semantic meanings. In this paper, we assume similarity semantic meanings are independent, which means the probability of a query having one similarity semantic meaning does not affect the probability of this query carrying other similarity semantic meanings. With different training datasets for each similarity semantic, the query dispatcher can be trained for each separately.

As noted before, meta-path-based features often carry different semantic meanings and these features can be used to measure similarity between entities from different aspects. Similarity query dispatcher is also defined in the scope of meta-path-based feature space. We here employ a popular prediction model which is similar to logistic regression to indicate whether the given query is related to one specific similarity semantic meaning or not.

$$Pr(match = 1|q, S; \boldsymbol{\theta}) = \frac{e^z}{e^z + 1} \tag{4.7}$$

where $z = \Sigma_{f_i \in F'} \theta_i \cdot f_i + b$. $Pr(match = 1|q, S; \boldsymbol{\theta})$ is the probability that query q matches similarity semantic meaning S. f_i is meta-path-based similarity feature.

In order to learn query similarity semantic meaning matching model, we use logistic regression with L_2 regularization to estimate the optimal $\boldsymbol{\theta}$ given a training dataset $\mathcal{T}$.

$$\hat{\boldsymbol{\theta}} = argmin_{\boldsymbol{\theta}} \Sigma_{i=1}^n - \log Pr(match|q; S; \boldsymbol{\theta}) + \mu \Sigma_{j=0}^d \theta_j^2 \tag{4.8}$$

With this objective function defined in Equation 4.8, weights in the probability model can be easily estimated with a number of optimization methods. We use standard MLE (Maximum Likelihood Estimation) in our experiments to derive $\hat{\boldsymbol{\theta}}$ which maximizes the likelihood of all the training pairs.

With the similarity semantic matching model for each semantic meaning learned, given a query from user, the probability of semantic matching between the query and this specific semantic meaning can be calculated. We can either rank the probability of the given query matching semantic meanings and pick the top-1 semantic meaning and use the corresponding ranking model to solve the query. Or as discussed before, each semantic matching model can determine whether the given query is a match or not, and all matching ranking models can be used to answer the given query.

4.3 Unified User Guided Entity Similarity Search Framework

During the on-line query answering process, users provide similarity queries aiming to retrieve similar entities from a heterogeneous information network dataset. A query with user guidance will first be passed to the query dispatcher, in which different similarity semantics are represented by a number of meta-path-based measurements, and the learned threshold of each will help the search framework measure the relevance between the query and the semantic meaning. Using these will allow the query dispatcher to find all related semantic meanings for the input query. For each matched semantic, a linear ranking model which is represented using meta-path-based features can be found in the ranking model ensemble set. Each similarity ranking model will take in the similarity query, rank all possible entity candidates of the same type in the information network, and return the final results separately.

5. EXPERIMENTS

We test the proposed semantic meaning aware user guided entity similarity search approach in this section. In order to demonstrate the existence of query semantic ambiguity problem, and the power of the proposed approach in terms of understanding similarity semantic meanings and retrieve higher relevance entities, we apply our method along with several competitor methods on both DBLP network and IMDb network. The schema of the two networks can be found in Figure 2. Both datasets are collected from live web-based applications. DBLP network contains $5,000$ authors, 464 publication venues (both conferences and journals), $382,519$ papers, $16,798$ terms, and IMDb network has $29,360$ movies and TV Shows, $53,088$ actors, $5,408$ directors, $22,470$ keywords and 28 different genres. Considering four entity types, which are author, publication venue, movie and actor, we define 17 different similarity search semantic meanings on these two networks. By utilizing additional similar examples provided by the user (as in the examples we introduced in Section 1), our approach can first infer entity similarity semantic meanings associated with given queries, and then choose corresponding ranking model(s) from ranking model ensemble to answer the query under the predicted semantic assumptions.

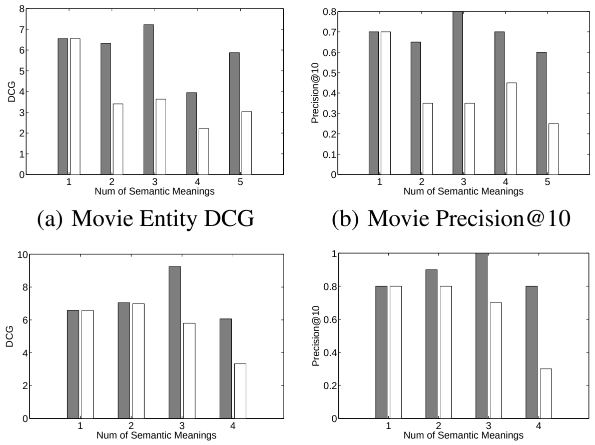
5.1 Training Dataset and Feature Space

We generate training datasets from previously published data mining results in literatures, by using well-known ground truth from public web services including DBLP and arnetminer etc, or by manually labeling. The quality and quantity of our training datasets are arguably sufficient for all the experiments but we do ensure the fairness of comparison by using the same training datasets when learning raking models for different systems.

In order to interpret various entity search semantic meanings, a comprehensive meta-path-based feature space needs to be calculated before model learning. We enumerate all valid meta-paths within length 6 between query entity types and use this set as $\mathbf{P}$. We implement PathSim, Personalized PageRank (denoted as p-PageRank), Random Walk, as well as SimRank as meta-path compatible measurements and use this set as $\mathbf{M}$. One should notice that, due to the differences in measurement definitions, similarity measurements have different meta-path compatibility. For instance, PathSim requires symmetric meta-paths, and p-PageRank can only be calculate along infinite paths (or paths which are long enough to make p-PageRank converge). What's more, meta-paths can be semantically insignificant, and measurements calculated along such paths can hardly contribute anything to the model learning process. When we calculate meta-path-based feature space $\mathbf{F}$, we ignore invalid meta-path and measurement combinations and skip semantically insignificant meta-paths for the consideration of efficiency.

5.2 Semantic Meaning Awareness Experiment

We define 17 different similarity semantic meanings over the entire dataset, and collect training datasets accordingly. In order to demonstrate the similarity semantic ambiguity problem, we incrementally add training datasets associated with different similarity semantic meanings into our system and the competitor system. Both systems utilize the entire meta-path-based feature space, use same optimization method and objective function during ranking model learning, and take in the same amount of training data at each stage. The difference is at each stage, a new ranking model is trained and similarity semantic meaning dispatcher is updated accordingly in our system, while in the competitor system, semantic meaning unaware learning-to-rank approach, combines new data and old data and learns one updated ranking model each time. We invite people with appropriate domain background knowledge to test our system, to assess the top 30 results with their judgment and with evidence. During evaluation process, they need to make relevance judgment for each returned entity of the top 30 results, and they need to provide evidence (using Google, or written details) of his / her judgment.



(a) Movie Entity DCG

(b) Movie Precision@10

(c) Author Entity DCG

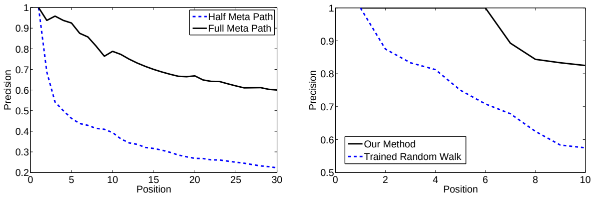
(d) Author Precision@10

Figure 3: Performance Plots with DCG and Precision@10

We use Discounted Cumulative Gain (DCG score), which is a widely used information retrieval measure considering both precision and recall, and precision-at-10 to evaluate the performances of the two systems and the results can be found in Figure 3.

During the experiments, to make a fair competition, we make sure that queries and related entities do not appear in any training datasets. Based on the experiments results, one can tell that, when there is only one semantic meaning in the training dataset, the performance of both systems are exactly the same, the newly proposed semantic meaning aware framework can retrieve more relevant results than the competitor system when the number of semantic meanings increases in the entity collection. Considering movies, even with only two semantic meanings, our system can increase the performance of regular learning-to-rank model by at least 5% measuring with precision-at-10. Significant improvements can be observed in other entity types as well. However, the improvement of our system is not linear with the number of semantic meanings increase, and little improvement is seen on publication venue and author entities when two semantic meanings are introduced in the system. The explanation of such observation is semantic meanings for one entity type may not be orthogonal, and semantic meanings of the same entity type are correlated somehow. When the correlation of semantic meanings in the system is high, improvement of our system will be less obvious.

This experiment proves the existence of the similarity semantic ambiguity problem in real world heterogeneous information network datasets, showcases the ability of entity similarity query semantic understanding of the proposed system, and also demonstrates the importance of semantic meaning understanding in terms of improving entity retrieval results.



(a) Actor Entity Precision (b) Publication Venue Prec.

Figure 4: Performance Plots of Different Meta-Path-Based Models

5.3 Expressiveness of Meta-Path-based Feature Space

A relatively comprehensive meta-path-based feature space should be more expressive in terms of representing different similarity semantic meanings than an incomplete or sampled similarity feature space. In these experiments, entity similarity semantic meanings are fixed and known to all comparison methods. We will change either $\mathbf{P}$ or $\mathbf{M}$ to reduce the completeness of feature space $\mathbf{F}$, and test the effect on the results and overall performance.

We first sample 50% meta-path set $\mathbf{P}$ several times, and use the sampled meta-path sets to build similarity feature space, and apply the exact same learning-to-rank technique with the same amount of training datasets. We apply both our proposed approach with the sampled meta-path-based ranking method on IMDb network, and measure precision at each position in top 30 results of both methods, and the results can be found in Figure 4(a). From the results, one can notice that our proposed ranking system which utilizes the entire meta-path set $\mathbf{P}$ outperforms the ranking model with half sampled meta-path set, by around 50% in actor queries. The reason a more comprehensive meta-path set can lead to better performance is that it can interpret more similarity semantic meanings than sampled meta-path set.

We then sample meta-path compatible measurement set $\mathbf{M}$, and we only use random walk as entity similarity measurement when building similarity feature space. We apply both the approach using the entire measurement set and the ranking model with only random walk similarity features on DBLP dataset, and similarly measure the precision at each position in top 30 results returned by both methods. The results can be found in Figure 4(b). Very similar to the previous experiments, by using more meta-path compatible similarity measurements, our method outperforms the ranking model which only uses random walk as similarity measurements. This means the meta-path-based similarity feature space is more expressive and can convey more information than traditional supervised random walk method. By aggregating different meta paths and a number of similarity measurements, meta-path-based feature space makes representing entity similarity from different semantic aspect possible, which is the prerequisite of high quality ranking model ensemble and query dispatcher.

6. CONCLUSION

In this paper, we address the problem of similarity query semantic meaning understanding and query processing in heterogeneous information networks. Meta-path-based feature space is defined in order to measure similarity from different aspects, which could be used in information network retrieval as well as other information network related applications. By learning ranking models for different similarity semantic meanings and training a query dispatcher which can infer the hidden similarity semantic meanings of a given query, our search framework first predict the hidden motivation and the semantic meaning in a given query based on the user's guidance, and then uses related ranking models to further answer the query by returning lists of related entities.

Empirical study shows that our approach describes and interprets different semantic meanings better than other meta-path-based methods; in addition, the significance of the semantic meaning understanding problem and the effectiveness of our method in terms of predicting semantic meanings based on a user's query in the meta-path-based feature space is demonstrated. Interesting future work includes, retrieval and semantic meaning prediction on multi-sourced heterogeneous information network, e.g., cyber-physical social network, meta-path-based feature space scalability study, and semantic meaning prediction on multi-typed queries.

7. ACKNOWLEDGEMENTS

The work was supported in part by ARL W911NF-09-2-0053 (NSCTA), NSF IIS-0905215, CNS-0931975, AFOSR MURI award FA9550-08-1-0265, and DHS.

8. REFERENCES

[1] H. Abdi. The kendall rank correlation coefficient. *Encyclopedia of Measurement and Statistics. Thousand Oaks (CA): Sage*, pages 1–7, 2007.

[2] S. Chakrabarti. Dynamic personalized pagerank in entity-relation graphs. In *WWW'07*, pages 571–580, 2007.

[3] C. Chang, Y. Du, J. Wang, S. Guo, and P. Thouin. Survey and comparative analysis of entropy and relative entropy thresholding techniques. In *Vision, Image and Signal Processing, IEE Proceedings-*, volume 153, pages 837–850. IET, 2006.

[4] X. Geng, T. Liu, T. Qin, and H. Li. Feature selection for ranking. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 407–414. ACM, 2007.

[5] S. Gu, J. Yan, L. Ji, S. Yan, J. Huang, N. Liu, Y. Chen, and Z. Chen. Cross domain random walk for query intent pattern mining from search engine log. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, pages 221–230. IEEE, 2011.

[6] G. Jeh and J. Widom. Simrank: a measure of structural-context similarity. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 538–543. ACM, 2002.

[7] N. Lao and W. Cohen. Relational retrieval using a combination of path-constrained random walks. *Machine learning*, 81(1):53–67, 2010.

[8] Y. Sun, R. Barber, M. Gupta, C. Aggarwal, and J. Han. Co-Author Relationship Prediction in Heterogeneous Bibliographic Networks. In *Proceedings of 2011 Int. Conf. on Advances in Social Network Analysis and Mining*. IEEE, 2011.

[9] Y. Sun, J. Han, X. Yan, S. P. Yu, and T. Wu. PathSim: Meta Path-Based Top-K Similarity Search in Heterogeneous Information Networks. In *Proceedings of the 37th International Conference on Very Large Data Bases*. ACM, 2011.

[10] X. Yu, Q. Gu, M. Zhou, and J. Han. Citation prediction in heterogeneous bibliographic networks. In *Proc. of Siam International Conference on Data Mining*, 2012.

[11] X. Yu, A. Pan, L. Tang, Z. Li, and J. Han. Geo-friends recommendation in gps-based cyber-physical social network. In *2011 International Conference on Advances in Social Networks Analysis and Mining*, pages 361–368. IEEE, 2011.