Entity Information Management in Complex Networks

Yi Fang
Department of Computer Science
250 N. University Street
Purdue University, West Lafayette, IN 47906, USA
fangy@cs.purdue.edu

ABSTRACT

Entity information management (EIM) is a nascent IR research area that investigates the information management process about entities instead of documents. It is motivated by the increasingly sophisticated user information needs that go beyond document search. In the recent years, entity retrieval especially expert search has attracted much attention in the IR community while many other EIM problems have been rarely investigated. On the other hand, the entities in the real world or in the Web environment often present a network structure among them, i.e., there exist explicit interactions or implicit dependencies among the related entities. Recently, the emergence of social media such as Facebook and Twitter has further exemplified this network structure of entities (e.g., users registered at these sites can become friends, a fan or a follower of others). The resulting networks are very complex in the sense that they are heterogeneous, large-scale, multi-lingual and dynamic. These complex networks go beyond traditional social network analysis.

In this proposed research, I investigate entity information management in the environment of complex networks. The main research question is: how can the EIM tasks be facilitated by modeling the content and structure of complex networks? The research is an intersection of content based information retrieval and complex network analysis, which deals with both unstructured text data and structured networks. The specific targeting EIM tasks are entity retrieval, entity profiling and entity distillation. The research methodology and proposed experiments are also discussed in this proposal.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.4 [**Information Systems Applications**]: H.4.2 Types of Systems; H.4.m Miscellaneous

General Terms

Algorithms, Measurement, Performance, Experimentation

Keywords

Entity retrieval, Entity profiling, Social networks

1. INTRODUCTION

Entity information management (EIM) deals with organizing, processing and delivering information about entities. Its emergence is a result of satisfying more sophisticated information needs that go beyond document search. As the Web has evolved into a data-rich repository, both commercial systems and the information retrieval community have shown increasing interest in not just returning documents, but specific objects or entities in response to a user's query. Many entity search engines have recently emerged to identify specific types of entities of interest such as people, restaurants, products and so on. TREC launched a new Entity track in 2009 to investigate the problem of related entity finding. In its pilot task, given the name and homepage of an entity, as well as a context described in natural language text, the retrieval system needs to find target entities with homepages that are of the specified type. The Entity track will continue in 2010 to further investigate the entity retrieval problem. INEX (INitiative for the Evaluation of XML Retrieval) has also started the XML Entity Ranking track[1] since 2007. Some EIM problems go beyond retrieval and ranking such as: 1) entity profiling, which is about characterizing a specific entity, and 2) entity distillation, which is about discovering the trend about the entity. These problems have been rarely researched while they have many important applications.

On the other hand, the entities in the real world or in the Web environment are usually not isolated. They are connected and related with each other in one way or another. In some cases, there exist explicit semantic relationships among the entities. For example, the coauthorship makes the authors with similar research interests be connected. The emergence of social media such as Facebook, Twitter and Youtube has further interweaved the related entities in a much larger scale. Millions of users in these sites can become friends, fans or followers of others, or taggers or commenters of different types of entities (e.g., bookmarks, photos and videos). There also exist implicit dependencies that cannot be described by semantic relations. The Case A

[1] http://www.inex.otago.ac.nz/tracks/entity-ranking/guidelines.asp

and Case B dependencies in [4] are the two examples. These explicit relationships and implicit dependencies can yield complex networks of related entities. These networks are complex in the sense that they are heterogeneous with multiple types of entities and of interactions, they are large-scale, they are multi-lingual, and they are dynamic and evolving constantly. These features of the complex networks go beyond traditional social network analysis and require further research.

In this proposed research, I investigate how EIM in the real world can be facilitated by modeling the content and the structure of complex networks. In other words, the research deals with both unstructured text data and structured networks and is an intersection of content based information retrieval and complex network analysis. The proposed research methodology is based on probabilistic models as they are very flexible and powerful in modeling the uncertainties in the networks. In my previous work, I have done the following related work along with my collaborators: 1) Discriminative graphical models are proposed in [4] to jointly discover faculty homepages by inference on the homepage dependence graph/network; 2) Mixture models are proposed in [6] to learn flexible combination strategies to rank experts in heterogeneous information sources; 3) The dependence of table elements is exploited in [8] to collectively perform the entity retrieval task. These works have shown the benefits of utilizing the dependencies of related entities and the power of probabilistic models for entity search. I will further develop probabilistic models to address the research questions raised in this proposal. Three specific EIM tasks, namely entity retrieval, entity profiling and entity distillation, are investigated by the proposed research methodology.

The next section discusses related work. Section 3 introduces the motivation for the proposed research. Section 4 describes the research tasks and questions. In section 5, the proposed research methodology is presented. Section 6 provides several options of testbeds for the evaluation of the proposed methodology. Section 7 concludes. In addition, the appendix is placed at the last page.

2. MOTIVATION

The proposed research is mainly motivated by the fact of increasing interactions among related entities of online communities. A lot of valuable work has been done in social network analysis by different disciplines such as sociology, computer science and economics. When it comes to entity information management, the complex networks expose some new features as follows:

- Heterogeneous: The complex networks usually involve multiple types of entities or interactions. In addition, entities at the same network can interact with each other in various forms, leading to heterogeneous types of interactions between them.

- Large-scale: Traditional social network analysis relies on manual surveys to collect interaction information of human subjects. In contrary, the complex networks are often in a much larger scale with millions of entities and interactions in the networks which can be automatically collected by computational tools. This poses a challenge of scalability.

- Dynamic: The complex networks are dynamically evolving everyday as new entities join the network and new connections occur between existing entities. This is especially evident in social medias and real time data sources such as Twitter. This dynamic effect is very interesting for some time-sensitive applications as people are often interested in the recent or updated information.

- Multi-lingual: In many cases, the complex networks are multi-lingual by nature (e.g. academic researcher network, developer networks, corporate knowledge bases, blogospheres, Web 2.0 portals).

Most existing research in expert search and entity retrieval in general ignores the network structure of entities and treat them independently. This may lead to unsatisfactory results especially in the novel setting of modern social media where the linked entities are heavily present. Intuitively, the network structure can help improve the performance of various entity information management tasks by utilizing collective intelligence of related entities. For example, we have shown in [4] that by considering three types of dependencies among candidate homepages, the accuracy of faculty homepage detection has been significantly improved. The three types of dependencies lead to a network of candidate homepages. In addition, relational properties do show their importance in many other domains. It can also help alleviate data missing problems because the related entities can now borrow information/data from each other. For example, a researcher's expertise can be inferred to some extent from his/her collaborators' expertise. A further question could be why one cannot reduce a heterogeneous network to several homogeneous ones for investigation. The reason is that the interaction information in one mode or one dimension might be too limited or sparse to be meaningful. It is helpful to utilize information from other sources for more effective entity information management.

2.1 A Motivating Example

An example of a complex network is faculty publications as illustrated in Figure 1. Different types of entities (faculty, conferences/journals, papers) exist in the network. Specifically, faculty are connected to papers through authorship. Faculty can also be directly connected to each other by links through their homepages. Papers are published at different venues (e.g., conferences, journals, workshops, thesis, etc). Papers connects papers by citations. Some faculty might relate to each other by serving simultaneously as journal editors or on conference program committees. In the network, there are multiple types of entities and entities relate to others (either the same type or different types) through different links. This is also called Multi-mode Networks in [12]. On the other hand, there are multiple types of interactions between the same set of entities. For example, a researcher can write a paper or cite a paper. A faculty member can be a PC member, a speaker, or just an observer of a conference. This type of networks is also called Multi-Dimensional Networks where each dimension of the network represents one facet of diverse interactions between entities [9].

These faculty publication networks can include as many as millions of entities (i.e., professors, students, papers, venues, etc). They evolve constantly as new researchers join in or researchers have new publications. The networks can be

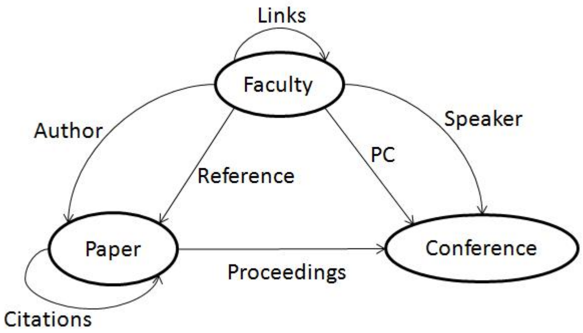

Figure 1: An example of complex networks in academia

multi-lingual especially when some researchers are not in the English speaking countries.

3. RESEARCH QUESTIONS AND TASKS

The general question guiding this proposal is this: How can the entity information management tasks be facilitated by modeling the content and the structure of complex networks? This research attempts to combine the content based information retrieval with complex/social network analysis. First of all, any EIM task involves building entity profiles by analyzing the content of online data which is usually in the form of unstructured text. Secondly, the EIM tasks are studies in the context of complex networks, since the network structure can provide valuable information about related entities. In sum, the proposed research exploit both unstructured data and structured networks, which is a novel research direction in EIM.

In particular, I address the main research question by investigating three specific EIM tasks: entity retrieval, entity profiling and entity distillation. The definitions of the three tasks are as follows.

- Entity retrieval: Given a topic, find relevant entities[2]. This is also the related entity finding task in TREC 2009.
- Entity profiling: Given an entity, find a list of key attributes of the entity. An example is to return a list of topics an expert is knowledgeable about.
- Entity distillation: Given an entity, find the currently or potentially important information about the entity.

These three research tasks are related to each other. The first two tasks rely on estimating the entities profiles and as pointed out in [2], they are essentially two sides of the same coin. The third task adds the time dimension and needs to consider the dynamic effect of entity profiles. Table 1 gives an examples of the results returned for the tasks. Like the Entity track in TREC 2009, entities are defined by their homepages. The entity retrieval task requests the homepage of the target entities. For entity profiling, a list of expertise areas are returned in this example for Prof. ChengXiang Zhai. Entity distillation tries to find the research areas that he is currently working on or to infer his research interest for the near future. Table 2 shows another example for products. It is assumed that the tasks had been performed long before the engine problems of Toyota cars attracted a lot of public attention. The entity distillation task is expected to extract the currently or potentially key topics from the discussions in the car forums. As we can see, the task has important applications in reputation management and alert services.

Table 1: An example of results presented for the EIM tasks with the target entity type "people"

	ChengXiang Zhai
Homepage:	http://www.cs.uiuc.edu/homes/czhai/
Profile:	Language models, Relevance feedback, Smoothing
	Risk minimization, Active feedback, Mixture models
	NLP, Personalized search
Distillation:	Gene ontology, Topic modeling

Table 2: An example of results presented for the EIM tasks with the target entity type "product"

	Toyota	
Homepage:	http://www.toyota.com/	
Profile:	Headquarter:	Toyota City, Aichi, Japan
	Founder:	Kiichiro Toyoda
	Industry:	Automotive
Distillation:	Engine problems, Accelerator, Brake	

Besides the main research question, the following more specific questions are also considered:

- How can we simultaneously accomplish a EIM task for diverse entity types? For example, in entity profiling, a faculty member's profile is related with the profile of the conferences he attend. To jointly determine the profiles of the faculty and conferences by collective intelligence can be more robust and effective than the individual decisions on profiles.
- How to model the evolution of entity profiles and complex networks in a unified way? Once knowing the dynamics, we can infer the trend about the entity and mine possibly very important hidden patterns.
- How can some data mining related tasks such as community analysis and link prediction help entity information management?
- Will the use of multi-lingual evidence affect building entity profiles in complex networks? If yes, how can the existing cross-language IR work be leveraged?

4. RELATED WORK

Entity information management is a nascent IR field to satisfy user information needs seeking for entities. It has attracted increased attention in the IR community since the launch of expert search task in TREC 2005. With rich related work [2], expert search can be viewed as a subarea of entity search, while there exists few work for search for

[2] The definition of entity retrieval in some references is more general and is roughly equivalent to EIM defined in this proposal

Table 3: The extent to which the existing work is devoted to each EIM task. "Much" denotes there exists much related work and "Few" denotes there exists few related work

Task \ Entity	People	Organization	Product
Resolution	Much	Much	Much
Retrieval	Much	Few	Few
Profiling	Few	Few	Few
Distillation	Few	Few	Few

other types of entities such as organizations and products. TREC 2009 has generalized expert search by including these three types of entities to be searched [3]. The top run in the track was achieved by our group. The key ingredient of our entity search system is the special treatment of table and list data by considering the relations among them [8]. Two entity search tasks, list completion and entity ranking, were proposed in [1] and implemented for INEX Entity Ranking track in 2007, 2008 and 2009. A necessary step in many EIM tasks is often to recognize entities, which is called entity resolution or entity disambiguation. There are a lot of relevant work in the NLP community for entity resolution. In contrast, very few work has been done for entity profiling and entity distillation. The entity workshops in TREC 2008 and TREC 2009 briefly discussed these two tasks[3], but they have not been included in the TREC Entity track yet. Table 3 summarizes relatively how much work has been done for each EIM task.

The proposed research is also related to cross-language information retrieval [10] and social network analysis, as EIM is performed in the complex networks with multi-lingual users. Social network analysis has rich literature since it is across various disciplines. In recent years, the data mining community has paid significant amount of attention in this area, as the network structure becomes omnipresent in the online community and it requires efficient computational methods to handle the huge volume of the data. Many interesting network problems were proposed [12, 9], but they focused on the structure of networks and largely ignored the content of the networks. Furthermore, their solutions were based on matrix manipulations.

5. RESEARCH METHODOLOGIES

I propose to use probabilistic models to address the above research questions. The complex networks involve a huge amount of uncertainties and probabilistic models can provide principled ways to accommodate them. Specifically, the following four probabilistic methods serve as the building blocks of my research methodologies.

- Discriminative graphical models. They are used to define a joint probabilistic model for a collection of related entities. Rather than do the work for each entity separately, discriminative graphical models provide a form of collective intelligence, where we simultaneously decide on all of the entities together, and therefore can explicitly take advantage of the relations between the related entities. Figure 2 shows an example of discriminative graphic model representation in [4] for a homepage dependence network. Three types of dependencies are considered in this model. Furthermore, discriminative models have been shown to outperform their generative models in two benchmark TREC datasets [7].

- Topic models. A topic model is a generative model for documents: it specifies a simple probabilistic procedure by which documents can be generated [11]. To make a new document, one chooses a distribution over topics. Then, for each word in that document, one chooses a topic at random according to this distribution, and draws a word from that topic. Standard statistical techniques can be used to invert this process, inferring the set of topics that were responsible for generating a collection of documents. Topic models are a powerful tool to do the content based information retrieval and analysis.

- Mixture modeling. In fact, topic models are also a class of mixture models in which documents are mixtures of topics. On the other hand, with their flexible modeling capability for unlabeled data, mixture models can go beyond topic models. For example, in the academic network shown in Figure 1, it may be only known that a faculty member is associated with a conference according to the extracted information, but not known whether he/she is a PC or a speaker. In this case, mixture models are suitable to model the membership of the faculty member based on his/her characteristics or features. We have illustrated an application of mixture models for ranking experts [6].

- Time series analysis. It is often associated with the discovery and use of patterns, and prediction of future values. It can help understand how the patterns of related entities and interactions are likely to evolve over time. In fact, similar to network analysis, one characteristic of time series analysis is that the data are not generated independently, but depend on the previous values. There have been extensive works in time series analysis as the time effect is ubiquitous in many disciplines. Dynamic probabilistic models such as dynamic Bayesian network have shown their effectiveness in modeling the uncertainties in the evolved networks.

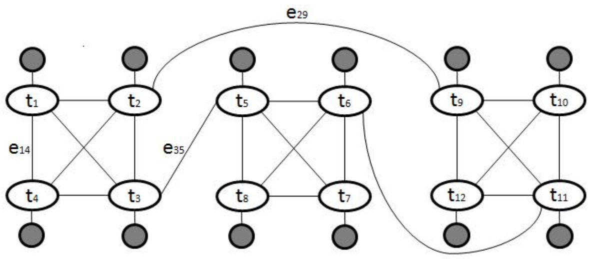

Figure 2: An example of discriminative graphic models in [4] for a homepage dependence network

[3] http://ilps.science.uva.nl/trec-entity/2008/12/workshop-summary/#more-9
http://ilps.science.uva.nl/trec-entity/guidelines/plans-for-2010/

As we can see, discriminative graphical models can capture network structures, topics models can analyze the unstructured text data, mixture models are able to handle uncertainties and missing data, time-series can capture the dynamics of the related entities. In consequence, the combinations of these probabilistic methods can yield very powerful tools for addressing the proposed research questions. [13] is an example of a probabilistic model of combining mixture modeling and time-series for complex networks (i.e., dynamic email communication network and gene interaction network).

6. PROPOSED EVALUATIONS

To evaluate the proposed research, the testbeds need to satisfy the properties of the complex networks defined in Section 2. The current TREC expert search datasets are not able to meet the requirement. Instead, the following testbeds can be good options.

- INDURE [5]. It is a research expertise database we developed at Purdue University. The whole system currently includes over 17,000 faculty members over four universities in the state of Indiana. The INDURE scenarios is very similar to that shown in Figure 1. In fact, some of the proposed research is motivated by the observations of the INDURE data. Another dataset with similar characteristics is the UvT Expert Collection[4]. It is bilingual (English and Dutch).

- DBLP. The DBLP data is constructed from the DBLP collection provided by knowledge discovery lab at UMass[5]. There are fewer relationships among entities in DBLP than in INDURE, but it is still a complex network.

- Twitter. It is the data used in 2010 for the WePS-3 Online Reputation Management Task[6]. The motivation is to help experts in reputation management and alert services, which coincides with the application of entity distillation. The test and training data consist of 500 names and 700 tweets for each name. The 700 tweets per name will be in English, Spanish or both.

- CriES Pilot Challenge. The pilot challenge of the CriES Workshop in 2010 instantiates the problem of expert search in multi-lingual social media[7]. Some of the goals and motivations of the CriES Workshop are consistent with this proposal, which to some extent validates my proposed research.

7. CONCLUSION

In summary, complex networks have become pervasive in the modern online community. They offer much valuable interaction information that can be exploited to enhance IR systems, but at the same time pose many great challenges for the IR research community. By the work of combining the network structure analysis with content based information retrieval, I genuinely expect that this line of research can help expand entity related applications and help better understand the complex networks that may have great social implications.

8. REFERENCES

[1] S. Adafre, M. de Rijke, and E. Sang. Entity retrieval. In *Proceedings of International Conference on Recent Advances in Natural Language Processing (RANLP-2007)*, 2007.

[2] K. Balog. People search in the enterprise. In *PhD thesis, University of Amsterdam*, 2008.

[3] K. Balog, A. de Vries, P. Serdyukov, P. Thomas, and T. Westerveld. Overview of the trec 2009 entity track. In *TREC-18*, 2009.

[4] Y. Fang, L. Si, and A. Mathur. Discriminative graphical models for faculty homepage discovery. *Information Retrieval*. `http://www.springerlink.com/content/vn8l56k53glv1360/`.

[5] Y. Fang, L. Si, and A. Mathur. FacFinder: search for expertise in academic institutions. Technical Report: SERC-TR-294, Department of Computer Science, Purdue University, 2008.

[6] Y. Fang, L. Si, and A. Mathur. Ranking experts with discriminative probabilistic models. In *Proceedings of SIGIR 2009 Workshop on Learning to Rank for Information Retrieval*, 2009.

[7] Y. Fang, L. Si, and A. Mathur. Discriminative Models of Integrating Document Evidence and Document-Candidate Associations for Expert Search. 33rd International ACM Conference on Research and Development in Information Retrieval (SIGIR), 2010.

[8] Y. Fang, L. Si, Z. Yu, Y. Xian, and Y. Xu. Entity retrieval by hierarchical relevance model, exploiting the structure of tables and learning homepage classifiers. In *TREC-18*, 2009.

[9] X. W. Lei Tang and H. Liu. Uncovering groups via heterogeneous interaction analysis. In *Proceedings of IEEE International Conference on Data Mining*, 2009.

[10] D. Oard and A. Diekema. Cross-language information retrieval. *Annual Review of Information Science and Technology (ARIST)*, 33:223–56, 1998.

[11] M. Steyvers and T. Griffiths. Probabilistic topic models. *Handbook of Latent Semantic Analysis*, pages 424–440, 2007.

[12] L. Tang, H. Liu, J. Zhang, and Z. Nazeri. Community evolution in dynamic multi-mode networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 677–685. ACM, 2008.

[13] F. W. Xing, E.P. and L. Song. A state-space mixed membership blockmodel for dynamic network tomography. *Annals of Applied Statistics*, 2009.

[4] http://ilk.uvt.nl/uvt-expert-collection/

[5] http://kdl.cs.umass.edu/data/dblp/dblp-info.html

[6] http://nlp.uned.es/weps/weps-3/guidelines/40-guidelines-for-the-weps-3-on-line-reputation-management-task

[7] http://www.multipla-project.org/cries:challenge