XIANG REN

RESEARCH STATEMENT

OVERVIEW

My research focuses on **turning massive unstructured text corpora into structured databases of factual knowledge**, for better management, exploration and analysis of large corpora. In today's computerized and information-based society, text data is rich but often also "messy". We are inundated with vast amounts of text corpora, written in different genres (from grammatical news articles and scientific papers to noisy social media posts), covering topics in various domains (e.g., medical records, corporate reports, legal acts). Can computational systems automatically identify various real-world entities mentioned in a new corpus and use them to summarize recent news events reliably? Can computational systems capture and represent different relations between biomedical entities from massive and rapidly emerging life science literature? How might computational systems represent the facts contained in a collection of medical reports as a relational table to support answering detailed queries or running data mining tasks?

While people can easily access the documents in a gigantic collection with the help of data management systems, they struggle to gain insights from such a large volume of text data: document understanding calls for in-depth content analysis, content analysis itself may require domain-specific knowledge, and over a large corpus, a complete read and analysis by domain experts will invariably be subjective, time-consuming and relatively costly. Moreover, text data is highly variable: corpora from different domains, genres or languages have typically required for effective processing a wide range of language resources (e.g., grammars, vocabularies, gazetteers). The "massive" and "messy" nature of text data poses significant challenges to creating tools for automated processing and algorithmic analysis of content that scale with text volume.

The goal of my research is to improve the mining of, and access to structured factual information reliably identified and extracted from unstructured text data, to overcome the barriers in dealing with text corpora of various domains, genres and languages. State-of-the-art information extraction (IE) systems have relied on large amounts of task-specific labeled data (e.g., annotating terrorist attack-related entities in web forum posts written in Arabic), to construct machine-learning models (e.g., deep neural networks). However, even though domain experts can manually create high-quality training data for specific tasks as needed, both the *scale* and *efficiency* of such a manual process are limited. My research harnesses the power of "big text data" and focuses on creating generic solutions for *efficient construction of customized machine-learning models for factual structure extraction*, relying on only limited amounts of (or even no) task-specific training data. The approaches that I developed are thus general and applicable to all kinds of text corpora in different natural languages, enabling quick deployment of data mining applications.

Research Summary. My dissertation opens up a new problem, called "*cold-start factual structure mining from massive text corpora*," (i.e., **cold-start StructMine**), where the full corpus analysis starts "cold", that is, with no prior manual annotation on the corpus. My research addresses the following question: *Is it possible to extract factual information from a corpus, and represent these inter-related facts using a unified structure that is machine-readable?* My work answers this question in depth—StructMine automates the process of extracting factual structures (which I define to include: entities of different categories [3], synonyms of entities [5], relationships of various relation types between entities [6]) from a large corpus *without human annotation data*, and integrates them with existing structures to construct a *structured information network* (henceforth **StructNet**). In contrast to knowledge graph approaches (e.g., Google Knowledge Vault, NELL, KnowItAll, DeepDive) that harvests facts incrementally from the whole Web to cover common knowledge in the world, my approach with StructNet provides a structured and unified view of all the facts in a given corpus, to enable semantic, holistic and scalable analysis of all content in the full corpus. Thus the construction of a corpus-specific StructNet is distinct from, but also complements the task of knowledge graph population. As a result, the application of cold-start StructMine techniques for building StructNets focuses on establishing only corpus-specific factual knowledge (e.g. identifying the entities and relations disambiguated for that corpus), a task that is outside the scope of general knowledge graphs.

Following the construction of a corpus-specific StructNet, my research addresses the question: *How might*

computational systems mine the StructNets to power text analysis applications? My approach leverages the extracted structure information already built into StructNets, where nodes that represent objects with different entity types are linked via edges that represent relationships with different relation types. I have developed computational methods that exploit the semantics of entity and relation types, to conduct *holistic* analysis on a StructNet and facilitate text analysis tasks (e.g., summarization [7], recommendation [8]).

My research solutions to the two questions provide a broadly conceived, two-phase architecture to analyze massive text corpora: (1) constructing StructNets from massive text corpora (i.e., **corpus-to-network**), and (2) mining the resulting StructNets to discover new insights and knowledge originating from the text corpora (i.e., **network-to-knowledge**). The impact and contributions of my research are summarized as follows.

- My work on cold-start StructMine has been awarded a **Google PhD Fellowship** in 2016; the findings of my research were presented as an essential part in the **ACL conference keynote** in 2015.
- Our systems for extracting typed entities and relations achieve over 25% improvement over state-of-the-art systems across various domains and genres [3, 1, 6]. The technology has been transferred to Microsoft Bing and Army Research Labs because of its excellent domain-independence and no reliance on human effort.
- Our phrase mining tool, SegPhrase [11], won the grand prize of Yelp Dataset Challenge in 2015, and has been shipped as parts of the products in Microsoft Bing and TripAdvisor.
- My tutorials on entity recognition and typing [4] attracted over 300 audience members in the top conferences of data mining, databases and information systems (KDD, SIGMOD, WWW).
- My work on StructNet mining received **Yahoo-DAIS Research Excellence Award** in 2015.

Research Philosophy. I select problems to work on whose solutions have the potential for direct impact on applications, and I apply an analytic approach to formulate these problems and an algorithmic approach to design and implement scalable solutions. Then, I conduct controlled experiments through real-world systems to test the effectiveness of the solutions. As a data mining researcher working in the domain of information extraction and NLP problems, my novel contributions to solving these problems have come from applying my expertise in machine learning and mathematical optimization solutions to these problems, by exploiting rich data redundancy (e.g., common text patterns) in massive text corpora and by leveraging existing natural language resources (e.g., Wikipedia) that overlap with the given corpora for distant supervision.

DISSERTATION RESEARCH

The first phase of my approach to making sense of a large text corpus is to find facts in the corpus, and to establish how these facts are inter-related: *What real-world entities and entity attributes are mentioned in the corpus? What kinds of relations are expressed between these entities in the corpus?* Corpus-specific StructNet construction is concerned with turning an unstructured text corpus into a StructNet, where nodes represent entities of different categories (e.g., `person`, `company`) with their attributes attached (e.g., `age`, `found_date`), and edges represent different types of relationships between these entities (e.g., `employee_of`).

Challenges. We have witnessed the great success of supervised machine-learning approaches in yielding state-of-the-art performance on factual structure extraction when abundant amounts of training data are available. In contrast to rule-based systems, supervised learning-based systems shift the human expertise in customizing systems away from the complex handcrafting of extraction rules to the annotation of training data and feature engineering. The resulting effectiveness of supervised IE systems largely depends on the amount of available annotated training data and complexity of the task. When the quantity of annotated data is limited and the complexity of the task is high, these factors become bottlenecks in development of supervised systems for cold-start StructMine tasks. Recent advances in bootstrapping pattern learning (e.g., NELL, KnowItAll, OpenIE) aim to reduce the amount of human involvement—only an initial set of annotated examples/patterns is required from domain experts, to iteratively produce more patterns and examples for the task. Such a process, however, still needs manual spot-checking of system intermediate output on a regular basis to avoid error propagation, and suffers from low coverage on "implicit relations", i.e., those that are not overtly expressed in the corpus and so fail to match textual patterns generated by the systems.

Proposed Solution.
My solution to cold-start StructMine problem aims to bridge the gap between customized machine-learning models and the absence of high-quality task-specific training data. It leverages the information overlap between background facts stored in external knowledge bases (KBs) (e.g., Wikipedia, BioPortal) and the given corpus to automatically generate large amounts of (possibly noisy) task-specific training data; and it exploits redundant text information within the massive corpus to reduce the complexity of feature generation (e.g., sentence parsing). This solution is based on two key intuitions which are described below.

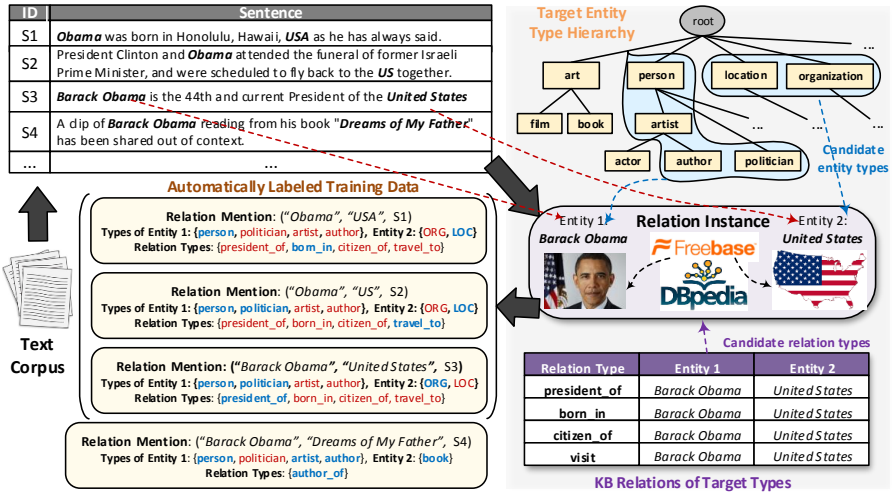

Figure 1. An illustration of distant supervision.

First, in a massive corpus, factual information about some of the entities (e.g., entity categories, relations to other entities) can be found in external KBs. My research asks, *can computational systems align the corpus with external KBs to automatically generate training data for building StructNet at a large scale?* For example, in Fig. 1, by mapping the word "*Obama*" in sentence S1 to the entity *Barack Obama* in Freebase, one can collect the entity category information about this entity (e.g., "*Obama*" could be a `person`, `politician`, etc.), as well as the relational facts about this entity (e.g., *Barack Obama* `is_president_of` *United States*). This retrieved information supports the automated annotation of entities and relations in text and labeling of their categories, yielding (possibly noisy) corpus-specific training data. Although the overlaps between external KBs and the corpus at hand might involve only a small proportion of the corpus, the scale of the automatically labeled training data could still be much larger than that of manually annotated data by domain experts.

Second, text units (e.g., word, phrase) co-occur frequently with entities and other text units in a massive corpus. My research asks, *can computational systems exploit the textual co-occurrence patterns to characterize the semantics of text units, entities, and entity relations?* For example, having observed that "*government*", "*speech*", "*party*" co-occur frequently with `politician` entities in the training data, the next time these text units occur together with an unseen entity in a sentence, the algorithm can more confidently guess that entity is a `politician`. As such patterns become more apparent in a massive corpus with rich data redundancy, big text data leads to big opportunities in representing semantics of text unit without complex feature generation.

A Principled Framework for Cold-Start StructMine.
To systematically model the intuitions above, we approach the cold-start StructMine tasks as follows: (1) annotate the text corpus automatically with target factual structure instances (e.g., entity names, entity categories, relations) by referring to external KBs, to create a task-specific training data (i.e., distant supervision); (2) extract shallow text units (e.g., words, n-grams, word shapes) surrounding the annotated instances in local context; (3) learn semantic vector representations for target instances, text units, and their category labels based on distant supervision and corpus-level co-occurrence statistics, through solving joint optimization problems; and (4) apply learned semantic vector representations to extract new factual instances in the remaining part of the corpus. The resulting framework, which integrate these ideas, has minimal reliance on human efforts, and thus can be ported to solve StructMine tasks on text corpora of different kinds (i.e., **domain-independent**, **language-independent**, **genre-independent**).

Method	NYT	Review	Tweet
Stanford NER	0.681	0.240	0.438
UW FIGER	0.881	0.198	0.308
ClusType [3]	**0.939**	**0.808**	**0.451**

Table 1. Compare ClusType with state-of-the-art entity recognition and typing systems on F1 scores.

To construct StructNet for a specific corpus, I apply the proposed framework to develop algorithms for three concrete subtasks, adopting a workflow as follows: (1) quality phrase mining [11], (2) "node mining" (i.e., entity recognition and typing [1, 2, 3]), and (3) "edge construction" (i.e., relation extraction and typing [6]).

First, we develop **SegPhrase** [11] algorithm to mine high-quality phrases from the corpus as candidate entity names, using entity names from external KBs as positive examples and corpus-level concordance statistics. SegPhrase demonstrates excellent domain independence compared with existing methods, and works well on low-resource languages (e.g., Arabic, Spanish). Second, we develop **ClusType** [3] algorithm to identify token spans in text that constitute entity mentions and to assign entity types to these spans. ClusType propagates type information from linkable mentions given by distant supervision to the remaining unlinkable ones, by leveraging relation phrases surrounding the entity mentions as the bridges. For instance, as shown in Table 1, ClusType system [3] achieves significant improvement over Stanford NER system and UW FIGER system on extracting entities for types of interest from NYT news articles, Yelp reviews and tweets. However, ClusType encounters several challenges in distinguishing fine-grained types with distant supervision: (1) KB types assigned to entity mentions might be noisy as the mapping is *context-agnostic*, and (2) *type correlation* in the given type hierarchy should not be ignored. I develop two novel techniques to resolve these challenges: (1) **PLE** [1] first de-noises the training instances, then learns models over the "cleaned" training data; and (2) **AFET** [2] jointly models the noisy types labels in the embedding process. Third, I develop **CoType** [6] algorithm for *joint* extraction of typed entities and relations with distant supervision. CoType embeds entity mentions, relation mentions and text units into low-dimensional vector space jointly; meanwhile, it also captures the cross-constraints of entities and relations on each other. Such a joint extraction process yields significant improvement on both entity typing and relation extraction tasks over the state-of-the-art systems [6] (e.g., see Table 2).

Method	NYT	BioInfer
IBM FCM	0.688	0.467
UW MultiR	0.693	0.501
CoType [6]	**0.851**	**0.617**

Table 2. Compare CoType with IBM neural network model and UW MultiR system on relation classification.

Applications of StructNet in Text Analytics. The proposed cold-start StructMine framework outlined above extracts high-quality building blocks for StructNets. By integrating these factual structures with existing structures associated with the corpus (e.g., meta data of documents, human interactions with documents), one can construct a unified StructNet to represent the entire corpus. A subsequent question is on how to mine these StructNets to support different text analysis applications.

Whereas existing studies on graph topological analysis can mine network links for knowledge discovery, they usually ignore the rich semantics of "type labels" on nodes and links. StructNet mining poses several unique challenges that should be addressed together: (1) how to perform holistic analysis over all the nodes and links in the network, (2) how to capture the semantic differences between various entity and relation types, and (3) how to model "human factors" associated with the network (e.g., user interactions with documents in the corpus).

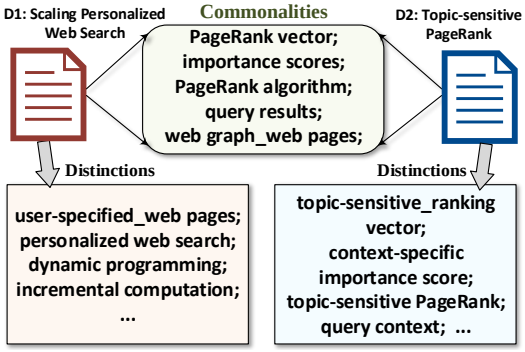

Figure 3. Comparative document analysis on two computer science papers.

In our solution, we formulate joint optimization problems to *collectively* model nodes and links in StructNet (i.e., holistic analysis), specify type-specific parameters in the models (i.e., type semantics), and personalize the models to capture user behavior differences (i.e., human factors). With these ideas in mind, I work on the problems of (1) generating informative yet concise summaries for text corpora [7, 10], (2) providing intelligent recommendations [8, 12], and (3) identifying search intents behind user queries [9]. My work answers questions such as: *What are the commonalities and distinctions between two sets of documents? What are the major events mentioned in a collection of news articles? What are the existing papers that an unpublished manuscript should cite?*

FUTURE RESEARCH AGENDA

My long-term goal is to create generic computational techniques to transform text data of various kinds into structured databases of human knowledge. I am passionate about the potential of data-driven thinking to improve the process of "structuring" knowledge contained in massive text data (generated by humans everyday and in everywhere)—so then powerful machines can act on the stored knowledge to improve human

productivity in various real-world application. I am also excited about applying our cold-start StructMine techniques to construct StructNets from scientific literature of various disciplines, such as life sciences, public health, social science, environmental science and economics—insights derived by analyzing the constructed StructNets can be used to benefit multi-disciplinary scientific research.

I plan to continue my research along the path of corpus-to-network-to-knowledge, to discover principles, propose methodologies, and design scalable solutions. The ideal outcome of my research is a generic system that can robustly construct StructNets from text corpora, regardless of the corpus nature, and can efficiently analyze the constructed StructNets for decision support and knowledge discovery. My mixed research background places me in a unique position for solving this challenging problem: my experience with natural language processing and text mining assists me in turning text data into StructNets, and my experience on machine learning and statistics enables me to develop scalable computational methods to analyze StructNets. The remainder of this section outlines some future opportunities that I am excited to pursue.

1. **Enrich Factual Representation.** In the current definition of StructNet, edges between two entities are weighted by the frequency of the facts mentioned in the text corpus. Such a representation has several limitations: (1) raw frequency cannot indicate *uncertainty* of the fact (e.g., drug A treats drug B with 75% success rate), (2) conditions of a relation are ignored in the modeling (e.g., if the patient is under 50 years old), and (3) complex relations involve more than two entities (e.g., protein `localization` relation). To address these challenges, I am interested in **collaborating with NLP researchers and linguists** to work on domain-independent sentiment analysis and syntax parsing for large text corpora, and incorporate the sophisticated linguistic features in StructNet construction. In particular, to measure fact uncertainty, it is critical to mine from a sentence words/phrases that indicate uncertainty (e.g., "*unlikely*", "*probably*", "*with 50% chance*"), negation (e.g., "*no*", "*barely*"), sentiments (e.g., "*efficiently*", "*nicely*"), or their enhancers (e.g., "very", "*extremely*"), and design systematic measures to quantify these units into weights of the edges in StructNets. To mine conditions for relations, I aim to extend the meta pattern-based attribute mining algorithm to identify patterns for "condition descriptions" (e.g., "*...[with age _]...*") and attach the mined conditions to edges in StructNet for further analysis. To extract complex relations, I plan to design scalable candidate generation process (e.g., different pruning strategy) to avoid producing exponential number of candidate relations, and extend the CoType embedding approach to model types for n-ary relations, while preserving the mutual constraints between relations and their entity arguments.

2. **Facilitate Scientific Research.** During scientific research, experts can only read a small subset of what is published in their fields, and are often unaware of developments in related fields. *Can we automate some parts of the research process to improve the productivity of researchers?* I am interested in: (1) **building next-generation literature search systems** for different research fields and disciplines, based on the StructNets constructed from the literature; (2) **forming candidate scientific hypotheses** (for further in-depth examination) by doing knowledge inference over the StructNets. On one hand, existing literature search systems are mostly keyword-based and thus have limited ability to satisfy complex information needs. StructNets index sentences and documents by typed entities and their facts—this provides great opportunities to work with domain experts to design new search (or question-answering) functions on the StructNet-indexed corpora (e.g., answering queries like "*What drugs can treat Alzheimer's disease?*"). On the other hand, by conducting collective inference on StructNets (e.g., link prediction), it should be possible to find useful linkages between information in related literatures (e.g., hypotheses on causes of diseases in biomedical domain), if the authors of those literatures rarely refer to one another's work.

3. **Engage with Human Behaviors and Interactions.** Structured human behavior data (e.g., social networks, user purchase records, web browser logs) helps construct better and richer StructNets, and conversely, StructNets (constructed from user-generated content) can **facilitate HCI systems and applications in computational social science**. I am interesting in working with researchers in HCI and social science to study challenging problems which arise when fusing StructNets with human behaviors and interactions. Such research questions include: *How to leverage social network structures to enrich text-based SturctNets? How to incorporate and analyze StructNets to help answer questions in political science, economics and psychology? How to build StructNet-powered HCI tools to improve user productivity?* To enrich text-

based StructNets, I plan to investigate principled methodologies for StructNet construction when social interactions between user-generated content are known—social interactions provide hints on correlation between documents and can potentially address data sparsity issues in the construction phase. To facilitate HCI systems, I plan to construct StructNets from text corpora describing user behaviors (e.g., fictions), and apply this database of human behavior facts (or inferred knowledge) to predict user's possible subsequent activities based on their current contexts (e.g., social media posts, photos taken).

4. **Integrate with the Physical World.** While StructNet construction relies on human-generated text content in cyber world, our physical world constantly produces data of various types that can be collected by physical sensors (e.g., geo-sensors in smart phones). I am interested in the **"fusion" of physical sensors and unstructured text data**—it can potentially improve the process of StructNet construction and help decision-makers better understand their physical environment. For instance, in smart city research, current software systems keep track of city-wide events and manage city-wide resources by deriving information exclusively from physical sensors. I plan to build a better system by accepting data not only from sensors but also from social media and current news. Such a system will incrementally update the StructNet for cyber world by relating text signals with sensor signals. The resulting fusion system can offer better StructNets (e.g., spatial-temporal information of a tweet helps disambiguate the entities mentioned in it) and situation understanding, leading to a better smart city operating system.

Collaboration and Funding. The research directions I intend to pursue require the expertise of researchers in many fields. I look forward to working closely with researchers who have expertise in natural language processing, databases, machine learning, systems, HCI, computer vision, and algorithms. Additionally, I plan to work with researchers outside computer science to identify typical usage patterns for text analysis in their field, be it life sciences, public health, or social science, and to determine whether text understanding and exploration can be profitably incorporated. I hope to bring to these collaborations connections to well-established principles underlying text analysis, and knowledge about how text analysis works in practice.
I was supported by NSF, ARL and Google PhD Fellowship, and will keep seeking funding opportunities in the future from multiple funding agencies (e.g., NSF, ARL, NIH, DARPA, ARO) and industries.

REFERENCES

[1] **Xiang Ren**, Wenqi He, Meng Qu, Heng Ji, Clare R. Voss, and Jiawei Han. Label noise reduction in entity typing by heterogeneous partial-label embedding. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (**KDD**)*, 2016.

[2] **Xiang Ren**, Wenqi He, Meng Qu, Heng Ji, and Jiawei Han. AFET: Automatic fine-grained entity typing by hierarchical partial-label embedding. In *Proc. of the Conf. on Empirical Methods in Natural Language Processing (**EMNLP**)*, 2016.

[3] **Xiang Ren**, Ahmed El-Kishky, Chi Wang, Fangbo Tao, Clare R. Voss, and Jiawei Han. ClusType: Effective entity recognition and typing by relation phrase-based clustering. In *Proceedings of the 21th ACM SIGKDD Intl. Conference on Knowledge Discovery and Data Mining (**KDD**)*, 2015.

[4] **Xiang Ren**, Ahmed El-Kishky, Heng Ji, and Jiawei Han. Automatic entity recognition and typing in massive text data (tutorial paper). In *Proceedings of the 2016 Intl. Conf. on Management of Data (**SIGMOD**)*, 2016.

[5] **Xiang Ren** and Tao Cheng. Synonym discovery for structured entities on heterogeneous graphs. In *Proceedings of the 24th Intl. Conference on World Wide Web (**WWW**)*, 2015.

[6] **Xiang Ren**, Zeqiu Wu, Wenqi He, Clare Voss, Heng Ji, Tarek Abdelzaher, and Jiawei Han. CoType: Joint extraction of typed entities and relations with knowledge bases. In *Proc.of the 26th Intl. Conf. on World Wide Web (**WWW**)*, 2017.

[7] **Xiang Ren**, Yuanhua Lv, Kuansan Wang, and Jiawei Han. Comparative document analysis for large text corpora. In *Proceedings of the 10th ACM Intl. Conference on Web Search and Data Mining (**WSDM**)*, 2017.

[8] **Xiang Ren**, Jialu Liu, Xiao Yu, Urvashi Khandelwal, Quanquan Gu, Lidan Wang, and Jiawei Han. ClusCite: effective citation recommendation by information network-based clustering. In *Proceedings of the 20th ACM SIGKDD Intl. Conference on Knowledge Discovery and Data Mining (**KDD**)*, 2014.

[9] **Xiang Ren**, Yujing Wang, Xiao Yu, Jun Yan, and Jiawei Han. Heterogeneous graph-based intent learning with queries, web pages and wikipedia concepts. In *Proc. of the ACM Intl. Conf. on Web Search and Data Mining (**WSDM**)*, 2014.

[10] Jialu Liu, **Xiang Ren**, Jingbo Shang, Taylor Cassidy, Clare R. Voss, and Jiawei Han. Representing documents via latent keyphrase inference. In *Proceedings of the 25th Intl. Conf. on World Wide Web (**WWW**)*, 2016.

[11] Jialu Liu, Jingbo Shang, Chi Wang, **Xiang Ren**, and Jiawei Han. Mining quality phrases from massive text corpora. In *Proceedings of the 2015 ACM SIGMOD Intl. Conf. on Management of Data (**SIGMOD**)*, 2015.

[12] Xiao Yu, **Xiang Ren**, Yizhou Sun, Quanquan Gu, Bradley Sturt, Urvashi Khandelwal, Brandon Norick, and Jiawei Han. Personalized entity recommendation: a heterogeneous information network approach. In *Proceedings of the 7th ACM Intl. Conference on Web Search and Data Mining (**WSDM**)*, 2014.