

## **Xiaodan (Tom) Zhang**

*Email:* xiaodanright@gmail.com

*Homepage:* <http://www.pages.drexel.edu/~xz38>

### **CAREER HIGHLIGHTS**

- Develop methods and algorithms to analyze and interpret critical internal and external information sources that enable drug discovery and development in a pharmaceutical / biomedical research environment. Assist in assembling a knowledge engineering infrastructure and incorporate unstructured text data into its knowledge base.
- Published more than 20 peer-reviewed conference and journal papers including SIGKDD09, SIGIR08, SIGIR06, IEEE ICDM 06, IJCAI 07, IEEE TKDE, SIAM SDM08 and SDM06, approximately 90 citations by Feb 2011.
- Won Drexel Best Doctoral Dissertation Award 2009; won Excellence in Research Award on Drexel Graduate Student Day 2009; received student travel awards for SIAM SDM 2006 and SIGIR 2008.
- Was key designer and developer of a text mining toolkit—Dragon Toolkit, which is widely used in the academic fields of text mining and information retrieval for both biomedical fields and general domains.
- Was key designer, developer and team leader of the first healthcare informatics system for cancer prevention and education in Pennsylvania—the Pennsylvania Cancer Education Network (PCEN) project.
- Built a prototype Metabolite Ontology (MO) by integrating public chemical databases and in-house Endogenous Metabolite Database (EMDB) at Merck Research Laboratories through two summer internships in 2007 and 2008. Expanded the contents of the in-house database

### **EDUCATION**

- PhD in Information Science: 09/2004 – 06/2009
- College of Information Science and Technology, Drexel University, Philadelphia, PA, USA
  - Advisor: Dr. Xiaohua Tony Hu
- Master of Computer Engineering: 09/2000 - 07/2003
- Computer Science Department, Jinan University, Guangzhou, China
  - Advisor: Dr. Weiqi Luo
- BS in library and information science 09/1993 - 07/1997
- Information Management Department, Northeastern Normal University, Changchun, China

### **SKILLS**

Design text mining, information retrieval, machine learning algorithms

Languages: Java, VB, C++, XML, ColdFusion, JSP, ASP, Perl, HTML.  
OS: Windows, Unix and Linux;  
Database: MYSQL, MSSQL

## **RESEARCH EXPERIENCE**

### **Vertex Pharmaceuticals Inc. Text Mining Engine**

- Research and develop text mining components of a system to capture important biomedical news & scientific articles, create links to internal & external resources, and deliver to relevant vertex groups; apply these components to Vertex Information Portal (VIP)—a news, literature, clinical trial, etc. knowledge distribution system
- Research and develop text mining methods to automatically extract protein-protein extraction information from biomedical literature.
- Expert network visualization, network centrality analysis

### **Dragon Toolkit (text data mining)**

- The Dragon Toolkit is a Java-based development package for academic use in information retrieval (IR) and text mining (TM), including text classification, text clustering, text summarization, and topic modeling. It is tailored for researchers who work on large-scale IR and TM and prefer Java programming. Moreover, in contrast to Lucene and Lemur, it provides built-in supports for semantic-based IR and TM. The dragon toolkit seamlessly integrates a set of NLP tools, which enable the toolkit to index text collections with various representation schemes including words, phrases, ontology-based concepts and relationships. However, to minimize the learning time, we intentionally kept the package small and simple.
- <http://dragon.ischool.drexel.edu/> (ICTAI'07)

### **The integration of external/domain knowledge into text mining using graph-based methods (dissertation research)**

- The graph representation of a text collection combined with ontology or social linkage information can enhance traditional text data mining such as ranking important node sets, clustering, classification and summarization (SIGKIDD'09, SIGIR'08, DAWAK'07, IJDWM'08).

### **Model based and semantic based text clustering and classification**

- A document is often full of class-independent “general” words and short of class-specific “core” words, which leads to the difficulty of document clustering. We argue that both problems will be relieved after suitable smoothing of document models using context sensitive phrases and ontology concepts (SIAM SDM'08, IEEE ICDM'06, IJCAI'07 and DASFAA'07).

### **Community detection and Annotation**

- Communities appear to play an important role in the functional properties of complex networks. Being able to annotate the identified community structure in a biological network can help us to understand better the **structure and dynamics of biological systems**. Thus, we present an ontology-based mixture language model approach to annotate protein community. (CIKM'06 and BIBE'06).

### **Semantic based text mining on Undiscovered Public Knowledge (UPK)**

- Two complementary and non-interactive literature sets of articles, when they are considered together, can reveal useful information of scientific interest not apparent in either of the two sets alone. This is referred to as UPK. Our method replaces manual ad-hoc pruning by using semantic knowledge from the biomedical ontologies. (SIAM DM'06 and ISI'06).

## SELECTED PUBLICATIONS

### *Book Chapters*

**Zhang X.**, Jing L., Hu X., Ng M.K., Xia J., Zhou X.:, Medical Document Clustering Using Ontology-Based Term Similarity Measures. Strategic Advancements in Utilizing Data Mining and Warehousing Technologies 2010: 133-150, 2010

**Zhang X.**, Hu X., Xia J., Zhou X., Achananuparp P., A Graph-Based Biomedical Literature Clustering Approach Utilizing Term's Global and Local Importance Information. Strategic Advancements in Utilizing Data Mining and Warehousing Technologies 2010: 133-150, 2010

### *Peer Reviewed Journal Papers*

- Hu X., **Zhang X.**, Yoo I., Wang X., Feng J., *Mining Hidden Connections among Biomedical Concepts from Disjoint Biomedical Literature Sets through Semantic-based Association Rule*, International Journal of Intelligent System, 25(2): 207-223 (2010) (1 Citations)
- Hu X., Park, E.K., **Zhang X.**, Microarray Gene Cluster Identification and Annotation through Cluster Ensemble and EM based Informative Textual Summarization, IEEE Transactions on Information Technology in Biomedicine, Sept., 2009, Vol. 13, No. 5, pp832-840
- **Zhang X.**, Jing L., Hu X., Ng M., Xia J., Zhou X., *Medical Document Clustering Using Ontology Based Term Similarity Measures*, International Journal of Data Warehousing and Mining (IJDWM), vol. 4 No. 1 pp. 62-71, 2008 (2 Citations)
- **Zhang X.**, Hu X., Xia J., Zhou X., Achananuparp P. *Utilization of Global Ranking Information in Graph- based Biomedical Literature Clustering* has been accepted in the 9th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2007) (acceptance rate: 30%, 45/150): one of the 5 best papers in DaWak 07 and extended versions was published in the International Journal of Data Warehousing and Mining, vol. 4. No.4 pp.84-101, 2008 (2 Citations)
- Zhou X., Hu X., **Zhang X.**: Topic Signature Language Models for Ad hoc Retrieval. IEEE Trans. Knowl. Data Eng. 19(9): 1276-1287 (2007) (11 Citations)
- Zhou X., Hu X., Li G., Lin X., **Zhang X.**, "Relation-based Document Retrieval for Biomedical IR", Transactions on Computational Systems Biology, Volume 4, Page 112-128, 2006 (1 Citations)

### *Journal Paper Under Review*

- **Zhang X.** et al. *Utilizing Different Link Types to Enhance Document Clustering based on Markov Random Field Model with Relaxation Labeling*, submitted to IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans

### *Peer Reviewed Conference Papers*

- Hu X, **Zhang X.** et al. Exploiting Wikipedia as External Knowledge for Document

Clustering, accepted to be published in 15<sup>th</sup> ACM SIGKDD Conference on Knowledge discovery and data mining, Paris June28th-July 1<sup>st</sup>, 2009 (acceptance rate: 12%) (20 citations) (13 Citations)

- Zhou X., Achananuparp P., Park E.K., Hu X., **Zhang X.**: AskDragon: a redundancy-based factoid question answering system with lightweight local context analysis. JCDL 2009: 483-484
- **Zhang X.**, Hu X., Zhou X., *A Comparative Evaluation of Different Link Types on Enhancing Document Clustering*, accepted in 31st Annual International ACM SIGIR Conference on Research & Development on Information Retrieval (**SIGIR 2008**) , (acceptance rate: 17%, 85/496) (7 citations)
- Zhou, X., **Zhang, X.** and Hu, X., *Semantic Smoothing for Bayesian Text Classification with Small Training Data*, to appear in the 2008 SIAM International Conference on Data Mining (SDM2008), April 24-26, Atlanta, Georgia (27%, 77/282) (1 Citations)
- Lu C., **Zhang X.**, Park J., Hu X. and He T., *Web Clustering based on the Information of Sibling Pages*, in 2008 IEEE International Conference on Granular Computing (GrC 08)
- Achananuparp P. Zhou X., Hu X., **Zhang X.**, *Semantic Representation in Text Classification Using Topic Singature Mapping*, To appear in Proceedings of 2008 IEEE International Joint Conference on Neural Networks, June 1-6, Hong Kong
- Achananuparp P., Hu X., Zhou X., **Zhang X.**, *Utilizing Sentence Similarity and Question Type Similarity to Response to Similar Questions in Knowledge-Sharing Community*, in WWW 2008 Workshop on Question Answering on the Web, April 22, Beijing, China 2008. (5 Citations)
- Zhou X., Hu X., **Zhang X.**, *A Segment-based Hidden Markov Model for Real-Setting Pinyin-to-Chinese Conversion*, in the Proceedings of the ACM CIKM 2007, pp 1027-1030 (acceptance rate: 26%, 512 submission) (2 Citations)
- **Zhang X.**, Hu X., et al., Achanauparp P. *Utilization of Global Ranking Information in Graph- based Biomedical Literature Clustering* has been accepted in the 9th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2007) (acceptance rate: 30%, 45/150): one of the 5 best papers in DaWak 07
- Zhou X, **Zhang X.**, and Hu X, "Dragon Toolkit: Incorporating Auto-learned Semantic Knowledge into Large-Scale Text Retrieval and Mining," In proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI), October 29-31, 2007, Patras, Greece, 197-201 (7 Citations)
- **Zhang X.**,Jing L., Hu X., Ng M., Zhou X., *A Comparative Study of Ontology Based Term Similarity Measures on Document Clustering*, the proceeding of the 12th International conference on Database Systems for Advanced Applications (**DASFFA 2007**) (acceptance rate: 18.7%, 70/373) (13 Citations)
- Zhou X., **Zhang X.**, and Hu X., "Semantic Smoothing of Document Models for Agglomerative Clustering," in the Twentieth International Joint Conference on Artificial Intelligence (IJCAI 2007), Jan. 6-12, 2007, India,2922-2927 (15.7%, 212/1353) (3 Citations)
- **Zhang X.** , Zhou X., and Hu X., *Semantic Smoothing for Model-based Document Clustering*, the proceeding of the 2006 IEEE International Conference on Data Mining (IEEE ICDM06), Dec. 18-22, 2006, HongKong (short paper, 800 submissions, 10% for full paper, 10% for short paper) (6 citations)
- **Zhang X.** , Wu D., Zhou X. , Hu X., *A Language Modeling Text Mining Approach to the Annotation of Protein Community*, the proceeding of the 6th IEEE Symposium on Bioinformatics and Bioengineering (BIBE 06), Nov 16-18, Washington DC, US (38.8%,

38/98), pp12-19

- Hu X., **Zhang X.**, Zhou X., *Integration of Cluster Ensemble and EM based Text Mining for Microarray Gene Cluster Identification and Annotation*, the Proceedings of ACM 15th Conference on Information and Knowledge Management (ACM **CIKM 2006**), poster paper, (537 submissions, 15% acceptance rate for full papers, 10% acceptance rate for post papers)
- Xu X., **Zhang X.**, Hu X., Using Two-stage Concept-based Singular Value Decomposition Techniques as a Query Expansion Strategy, *accepted to be published in the 2007 IEEE International Symposium on Data Mining and Information Retrieval*
- Chen Wu, Hu X, Shen X, **Zhang X.**, Yi Pan, *An Incremental Algorithm for Mining Default Definite Decision Rules from Incomplete Decision Tables* accepted to be published in the 2007 IEEE International Conference on Granular Computing
- Zhou X., Hu X., **Zhang X.**, Lin X., Song I.-Y., "Context-Sensitive Semantic Smoothing for the Language Modeling Approach to Genomic IR", the 29th Annual International ACM SIGIR Conference (ACM SIGIR 2006), Aug 6-11, 2006, Seattle, WA, USA, Page 170-177 (18.5%, 74/399) (24 Citations)
- Zhou, X., **Zhang, X.**, Hu, X., *MaxMatcher: Biological Concept Extraction Using Approximate Dictionary Lookup*, the proceeding of the 9th biennial The Pacific Rim International Conference on Artificial Intelligence (**PRICAI 2006**), Aug 9-11, 2006, Guilin, Guangxi, China, Page 1145-1149 (short paper, 596 submissions, 14.1% for full papers, 16.8% for short papers) (5 citations)
- Hu X., **Zhang X.**, Yoo I, Zhang Y-Q, *A Semantic Approach for Mining Hidden Links from Complementary and Non-Interactive Biomedical Literature*, the proceeding of 2006 SIAM Conference on Data Mining (**SIAM DM 2006**), pp200-209
- Zhou X., **Zhang X.**, Hu X., Using Concept-based Indexing to Improve Language Modeling Approach to Genomic, in the proceedings of the 28th European Conference on Information Retrieval (ECIR 2006) , pp 444-455 , (acceptance rate: 20%, 37/178) (6 citations)
- Zhou X., Hu X., Lin X., Han H., **Zhang X.**, "Relation-based Document Retrieval for Biomedical Literature Databases", 11th International Conference on Database Systems for Advanced Applications (DASFAA 2006), April 12-15, Singapore, Page 689-701 (25%, 47/188) (4 Citations)
- Hu X., **Zhang X.**, D. Wu., X. Zhou, P. Rumm, *Integration of Instance-based learning & Text Mining for Identification of Potential Virus / Bacterium as Bio-terrorism Weapons*, the proceeding of 2006 IEEE Intelligence and Security Informatics Conference(**ISI 2006**) (short paper) (1 citations)
- Hu X., **Zhang X.**, Zhou X, Wu D., "Integration of Cluster Ensemble and Language Modeling based EM Informative Textual Summrization for Gene Expression Analysis", 2006 IEEE International Conference on Systems, Man, and Cybernetics, Oct 8-11, 2006, Taipei, Taiwan

## **WORK EXPERIENCE**

### **Vertex Pharmaceuticals Inc. 12/2009 -- Present**

- **Research Scientist I (Text Mining)**
  - **Biomedical Literature Text Mining for Drug Discovery**

### **LYZ Capital Advisors 01/2009 –11/ 2009**

- System Developer and Quantitative Trader
  - Responsibilities includes generating trade files, monitoring

portfolios and trades, reporting /analyzing daily PnL, attributions, trading costs, and other portfolio analytics, developing trading and operation systems, risk monitoring systems, and portfolio analytics systems.

**Jinan University, Guangzhou, China [1997-2004]**

- Automation and Reference librarian
- Lecturer of Library Services
- Visiting scholar to University of Wisconsin - Eau Claire [01/2002-08/2002]
  - Worked at reference desk
  - Developed an online reference system.

**TEACHING  
EXPERIENCE**

**Teaching Assistant for the course of Database Management Systems (Drexel University, USA). [Fall term of 2008 and Winter term of 2009]**

- Lectured
- Graded assignments and answered questions.

**Instructor for the elective course of Information Search (Jinan University, China) [1999-2004]**

- Taught students how to use electronic databases and library services

**INTERN  
EXPERIENCE**

**Merck Research Laboratory (West Point, PA, USA) [Summer in 2007 and 2008]**

- Metabolite Ontology Toolkit [06/11/2007 – 08/31/2007]
  - Built a prototype Metabolite Ontology (MO) by integrating public chemical databases
- EMDB database project and Pathway enrichment analysis project [06/23/2008-08/29/2008]
  - Expanded the contents of in-house Endogenous Metabolite Database (EMDB)

**PRESENTATIONS**

SIAM DM'06, BIBE'06, ICDM'06, BIBM'08

**AWARDS**

Drexel Best Doctoral Dissertation Award 2009 (one of the two Best Doctoral Dissertation awards selected from all the PhD students graduated in 2009)  
Excellence in Research Award on Drexel Graduate Student Day 2009  
Student travel award for SIAM DM 2006 and SIGIR 2008  
Drexel Research Day Award 2007

**VOLUNTEERS,  
PROFESSIONAL  
MEMBERSHIPS,  
AND SERVICES**

Reviewer of Third ACM International Conference on Web Search and Data Mining, 17<sup>th</sup> ACM International Conference on Information and Knowledge Management, Journal of Data and Knowledge Engineering, Journal of IEEE Transactions on Knowledge and Data Engineering, Journal of IEEE/ACM Transactions on Computational Biology and Bioinformatics, Journal of Pattern Recognition  
PC member of ACM CIKM'09  
Member of ACM, IEEE and Information System (IS)  
Registration Chair of IEEE BIBM 2007 Conference  
Editor of IEEE BIBM 2008 Workshop proceedings

Student Volunteers of BIBM 2008, 2007 and SIGKDD 2006.