Successful Data Mining Methods for NLP

Jiawei Han
Dept. of Computer Science
Univ. of Illinois at
Urbana-Champaign
Urbana, IL 61801, USA
hanj@cs.uiuc.edu

Heng Ji
Computer Science Dept.
Rensselaer Polytechnic
Institute
Troy, NY 12180, USA
jih@rpi.edu

Yizhou Sun
College of Computer and
Information Science
Northeastern University
Boston, MA 02115, USA
yzsun@ccs.neu.edu

1 Overview

Historically Natural Language Processing (NLP) focuses on unstructured data (speech and text) understanding while Data Mining (DM) mainly focuses on massive, structured or semi-structured datasets. The general research directions of these two fields also have followed different philosophies and principles. For example, NLP aims at deep understanding of individual words, phrases and sentences ("micro-level"), whereas DM aims to conduct a high-level understanding, discovery and synthesis of the most salient information from a large set of documents when working on text data ("macro-level"). But they share the same goal of distilling knowledge from data. In the past five years, these two areas have had intensive interactions and thus mutually enhanced each other through many successful text mining tasks. This positive progress mainly benefits from some innovative intermediate representations such as "heterogeneous information networks" [Han et al., 2010, Sun et al., 2012b].

However, successful collaborations between any two fields require substantial mutual understanding, patience and passion among researchers. Similar to the applications of machine learning techniques in NLP, there is usually a gap of at least several years between the creation of a new DM approach and its first successful application in NLP. More importantly, many DM approaches such as gSpan [Yan and Han, 2002] and RankClus [Sun et al., 2009a] have demonstrated their power on structured data. But they remain relatively unknown in the NLP community, even though there are many obvious potential applications. On the other hand, compared to DM, the NLP community has paid more attention to developing large-scale data annotations, resources, shared tasks which cover a wide range of multiple genres and multiple domains. NLP can also provide the basic building blocks for many DM tasks such as text cube construction [Tao et al., 2014]. Therefore in many scenarios, for the same approach the NLP experiment setting is often much closer to real-world applications than its DM counterpart.

We would like to share the experiences and lessons from our extensive inter-disciplinary collaborations in the past five years. The primary goal of this tutorial is to bridge the knowledge gap between these two fields and speed up the transition process. We will introduce two types of DM methods: (1). those state-of-the-art DM methods that have already been proven effective for NLP; and (2). some newly developed DM methods that we believe will fit into some specific NLP problems. In addition, we aim to suggest some new research directions in order to better marry these two areas and lead to more fruitful outcomes. The tutorial will thus be useful for researchers from both communities. We will try to provide a concise roadmap of recent perspectives and results, as well as point to the related DM software and resources, and NLP data sets that are available to both research communities.

2 Outline

We will focus on the following three perspectives.

2.1 Where do NLP and DM Meet

We will first pick up the tasks shown in Table 1 that have attracted interests from both NLP and DM, and give an overview of different solutions to these problems. We will compare their fundamental differences in terms of goals, theories, principles and methodologies.

Proceedings of the Tutorials of the 53rd Annual Meeting of the ACL and the 7th IJCNLP, pages 1–4,
Beijing, China, July 26-31, 2015. ©2015 Association for Computational Linguistics

Tasks	DM Methods	NLP Methods
Phrase mining / Chunking	Statistical pattern mining [El-Kishky et al., 2015; Danilevsky et al., 2014; Han et al., 2014]	Supervised chunking trained from Penn Treebank
Topic hierarchy / Taxonomy construction	Combine statistical pattern mining with information networks [Wang et al., 2014]	Lexical/Syntactic patterns (e.g., COLING2014 workshop on taxonomy construction)
Entity Linking	Graph alignment [Li et al., 2013]	TAC-KBP Entity Linking methods and Wikification
Relation discovery	Hierarchical clustering [Wang et al., 2012]	ACE relation extraction, bootstrapping
Sentiment Analysis	Pseudo-friendship network analysis [Deng et al., 2014]	Supervised methods based on linguistic resources

Table 1. Examples for Tasks Solved by Different NLP and DM Methods

2.2 Successful DM Methods Applied for NLP

Then we will focus on introducing a series of effective DM methods which have already been adopted for NLP applications. The most fruitful research line exploited Heterogeneous Information Networks [Tao et al., 2014; Sun et al., 2009ab, 2011, 2012ab, 2013, 2015]. For example, the meta-path concept and methodology [Sun et al., 2011] has been successfully used to address morph entity discovery and resolution [Huang et al., 2013] and Wikification [Huang et al., 2014]; the Co-HITS algorithm [Deng et al., 2009] was applied to solve multiple NLP problems including tweet ranking [Huang et al., 2012] and slot filling validation [Yu et al., 2014]. We will synthesize the important aspects learned from these successes.

2.3 New DM Methods Promising for NLP

Then we will introduce a wide range of new DM methods which we believe are promising to NLP. We will align the problems and solutions by categorizing their special characteristics from both the linguistic perspective and the mining perspective. One thread we will focus on is graph mining. We will recommend some effective graph pattern mining methods [Yan and Han, 2002&2003; Yan et al., 2008; Chen et al., 2010] and their potential applications in cross-document entity clustering and slot filling. Some recent DM methods can also be used to capture implicit textual cues which might be difficult to generalize using traditional syntactic analysis. For example, [Kim et al., 2011] developed a syntactic tree mining approach to predict authors from papers, which can be extended to more general stylistic analysis. We will carefully survey the major challenges and solutions that address these adoptions.

2.4 New Research Directions to Integrate NLP and DM

We will conclude with a discussion of some key new research directions to better integrate DM and NLP. What is the best framework for integration and joint inference? Is there an ideal common representation, or a layer between these two fields? Is Information Networks still the best intermediate step to accomplish the Language-to-Networks-to-Knowledge paradigm?

2.5 Resources

We will present an overview of related systems, demos, resources and data sets.

3 Tutorial Instructors

Jiawei Han is Abel Bliss Professor in the Department of Computer Science at the University of Illinois. He has been researching into data mining, information network analysis, and database systems, with over 600 publications. He served as the founding Editor-in-Chief of ACM Transactions on Knowledge Discovery from Data (TKDD). He has received ACM SIGKDD Innovation Award (2004), IEEE Computer Society Technical Achievement Award (2005), IEEE Computer Society W. Wallace McDowell Award (2009), and Daniel C. Drucker Eminent Faculty Award at UIUC (2011). He is a Fellow of ACM and a Fellow of IEEE. He is currently the Director of Information Network Academic Research Center (INARC) supported by the Network Science-Collaborative Technology Alliance (NS-CTA) program of U.S. Army Research Lab and also the Director of KnowEnG, an NIH Center of Excellence in big data computing as part of NIH Big Data to Knowledge (BD2K) initiative. His co-authored textbook "Data Mining: Concepts and Techniques" (Morgan Kaufmann) has been adopted worldwide. He has delivered tutorials in many reputed international conferences, including WWW'14, SIGMOD'14 and KDD'14.

Heng Ji is Edward H. Hamilton Development Chair Associate Professor in Computer Science Department of Rensselaer Polytechnic Institute. She received "AI's 10 to Watch" Award in 2013, NSF CAREER award in 2009, Google Research Awards in 2009 and 2014 and IBM Watson Faculty Awards in 2012 and 2014. In the past five years she has done extensive collaborations with Prof. Jiawei Han and Prof. Yizhou Sun on applying data mining techniques to NLP problems and jointly published 15 papers, including a "Best of SDM2013" paper and a "Best of ICDM2013" paper. She has delivered tutorials at COLING2012, ACL2014 and NLPCC2014.

Yizhou Sun is an assistant professor in the College of Computer and Information Science of Northeastern University. She received her Ph.D. in Computer Science from the University of Illinois at Urbana Champaign in 2012. Her principal research interest is in mining information and social networks, and more generally in data mining, database systems, statistics, machine learning, information retrieval, and network science, with a focus on modeling novel problems and proposing scalable algorithms for large scale, real-world applications. Yizhou has over 60 publications in books, journals, and major conferences. Tutorials based on her thesis work on mining heterogeneous information networks have been given in several premier conferences, including EDBT 2009, SIGMOD 2010, KDD 2010, ICDE 2012, VLDB 2012, and ASONAM 2012. She received 2012 ACM SIGKDD Best Student Paper Award, 2013 ACM SIGKDD Doctoral Dissertation Award, and 2013 Yahoo ACE (Academic Career Enhancement) Award.

Reference

[Chen et al., 2010] Chen Chen, Xifeng Yan, Feida Zhu, Jiawei Han, and Philip S. Yu. 2010. Graph OLAP: A Multi-Dimensional Framework for Graph Data Analysis. Knowledge and Information Systems (KAIS).

[Danilevsky et al., 2014] Marina Danilevsky, Chi Wang, Nihit Desai, Xiang Ren, Jingyi Guo, and Jiawei Han. 2014. Automatic Construction and Ranking of Topical Keyphrases on Collections of Short Documents. Proc. 2014 SIAM Int. Conf. on Data Mining (SDM'14).

[Deng et al., 2009] Hongbo Deng. Michael R. Lyu and Irwin King. 2009. A Generalized Co-HITS algorithm and its Application to Bipartite Graphs. Proc. KDD2009.

[Deng et al., 2014] Hongbo Deng, Jiawei Han, Hao Li, Heng Ji, Hongning Wang, and Yue Lu. 2014. Exploring and Inferring User-User Pseudo-Friendship for Sentiment Analysis with Heterogeneous Networks. Statistical Analysis and Data Mining, Feb. 2014.

[El-Kishky et al., 2015] Ahmed El-Kishky, Yanglei Song, Chi Wang, Clare R. Voss, Jiawei Han. 2015. Scalable Topical Phrase Mining from Text Corpora. Proc. PVLDB 8(3): 305 – 316.

[Han et al., 2010] Jiawei Han, Yizhou Sun, Xifeng Yan, and Philip S. Yu. 2010. Mining Heterogeneous Information Networks. Tutorial at the 2010 ACM SIGKDD Conf. on Knowledge Discovery and Data Mining (KDD'10), Washington, D.C., July 2010.

[Han et al., 2014] Jiawei Han, Chi Wang, Ahmed El-Kishky. 2014. Bringing Structure to Text: Mining Phrases, Entity Concepts, Topics, and Hierarchies. KDD2014 conference tutorial.

[Huang et al., 2013] Hongzhao Huang, Zhen Wen, Dian Yu, Heng Ji, Yizhou Sun, Jiawei Han and He Li. 2013. Resolving Entity Morphs in Censored Data. Proc. the 51st Annual Meeting of the Association for Computational Linguistics (ACL2013).

[Huang et al., 2014] Hongzhao Huang, Yunbo Cao, Xiaojiang Huang, Heng Ji and Chin-Yew Lin. 2014. Collective Tweet Wikification based on Semi-supervised Graph Regularization. Proc. the 52nd Annual Meeting of the Association for Computational Linguistics (ACL2014).

[Kim et al., 2011] Sangkyum Kim, Hyungsul Kim, Tim Weninger, Jiawei Han, Hyun Duk Kim, "Authorship Classification: A Discriminative Syntactic Tree Mining Approach", in Proc. of 2011 Int. ACM SIGIR Conf. on Research & Development in Information Retrieval (SIGIR'11), Beijing, China, July 2011.

[Li et al., 2013] Yang Li, Chi Wang, Fangqiu Han, Jiawei Han, Dan Roth, Xifeng Yan. 2013. Mining Evidences for Named Entity Disambiguation. Proc. of 2013 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'13). pp. 1070-1078.

[Sun et al., 2009a] Yizhou Sun, Jiawei Han, Peixiang Zhao, Zhijun Yin, Hong Chen and Tianyi Wu. 2009. RankClus: Integrating Clustering with Ranking for Heterogeneous Information Network Analysis. Proc. the 12th International Conference on Extending Database Technology: Advances in Database Technology.

[Sun et al., 2009b] Yizhou Sun, Yintao Yu, and Jiawei Han. 2009. Ranking-Based Clustering of Heterogeneous Information Networks with Star Network Schema. Proc. 2009 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'09).

[Sun et al., 2011] Yizhou Sun, Jiawei Han, Xifeng Yan, Philip S. Yu and Tianyi Wu. 2011. PathSim: Meta Path-Based Top-K Similarity Search in Heterogeneous Information Networks. Proc. International Conference on Very Large Data Bases (VLDB2011).

[Sun et al., 2012a] Yizhou Sun, Brandon Norick, Jiawei Han, Xifeng Yan, Philip S. Yu, and Xiao Yu. Integrating Meta-Path Selection with User Guided Object Clustering in Heterogeneous Information Networks. Proc. of 2012 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'12).

[Sun et al., 2012b] Yizhou Sun and Jiawei Han. 2012. Mining Heterogeneous Information Networks: Principles and Methodologies, Morgan & Claypool Publishers.

[Sun et al., 2013] Yizhou Sun, Brandon Norick, Jiawei Han, Xifeng Yan, Philip S. Yu, Xiao Yu. 2013. PathSelClus: Integrating Meta-Path Selection with User-Guided Object Clustering in Heterogeneous Information Networks. ACM Transactions on Knowledge Discovery from Data (TKDD), 7(3): 11.

[Sun et al., 2015] Yizhou Sun, Jie Tang, Jiawei Han, Cheng Chen, and Manish Gupta. 2015. Co-Evolution of Multi-Typed Objects in Dynamic Heterogeneous Information Networks. IEEE Trans. on Knowledge and Data Engineering.

[Tao et al., 2014] Fangbo Tao, Jiawei Han, Heng Ji, George Brova, Chi Wang, Brandon Norick, Ahmed El-Kishky, Jialu Liu, Xiang Ren, Yizhou Sun. 2014. NewsNetExplorer: Automatic Construction and Exploration of News Information Networks. Proc. of 2014 ACM SIGMOD Int. Conf. on Management of Data (SIGMOD'14).

[Wang et al., 2012] Chi Wang, Jiawei Han, Qi Li, Xiang Li, Wen-Pin Lin and Heng Ji. 2012. Learning Hierarchical Relationships among Partially Ordered Objects with Heterogeneous Attributes and Links. Proc. 2012 SIAM International Conference on Data Mining.

[Wang et al., 2014] Chi Wang, Jialu Liu, Nihit Desai, Marina Danilevsky, and Jiawei Han. 2014. Constructing Topical Hierarchies in Heterogeneous Information Networks. Proc. Knowledge and Information Systems (KAIS).

[Yan et al., 2008] Xifeng Yan, Hong Cheng, Jiawei Han, and Philip S. Yu. 2008. Mining Significant Graph Patterns by Scalable Leap Search. Proc. 2008 ACM SIGMOD Int. Conf. on Management of Data (SIGMOD'08).

[Yan and Han, 2002] Xifeng Yan and Jiawei Han. 2002. gSpan: Graph-Based Substructure Pattern Mining. Proc. 2002 of Int. Conf. on Data Mining (ICDM'02).

[Yan and Han, 2003] Xifeng Yan and Jiawei Han. 2003. CloseGraph: Mining Closed Frequent Graph Patterns. Proc. 2003 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'03), Washington, D.C., Aug. 2003.

[Yu et al., 2014] Dian Yu, Hongzhao Huang, Taylor Cassidy, Heng Ji, Chi Wang, Shi Zhi, Jiawei Han, Clare Voss and Malik Magdon-Ismail. 2014. The Wisdom of Minority: Unsupervised Slot Filling Validation based on Multi-dimensional Truth-Finding. Proc. The 25th International Conference on Computational Linguistics (COLING2014).