Heterogeneous Information Networks for Text-Based Link Mining: A Position Paper on Visualization and Structure Learning Methods

William H. Hsu
Dept. of Computing and Information Sciences
Kansas State University
bhsu@ksu.edu
+1 785 236 8247

Praveen Koduru
iQGateway, LLC
Praveen.Koduru@gmail.com

ChengXiang Zhai
Dept. of Computer Science
University of Illinois
czhai@illinois.edu

Abstract

In this position paper, we discuss the representation of user and community domains in blogs, forums, and social media as heterogeneous information networks, and describe a broad framework for analyzing interrelationships within such networks. This approach builds upon two main methodologies: one that is focused on community detection and graph analysis; and another that infers topic models in relation to communities and group-level topic mixtures. We first discuss previous work on link existence prediction in social networks, particularly user-touser links (*e.g.*, friend recommendation) and userto-community links (*e.g.*, community detection and recommendation), and point out some critical limitations of using this approach to detect communities. Next, we present a very general and flexible probabilistic model for topic modeling across heterogeneous information networks, and survey techniques for learning this

We then discuss why this approach has become necessary given changes in the privacy policies of social networks and other information service providers. Finally, we discuss the task of synthesize structure learning and topic modeling, and relate this to extant approaches, applications, and future work.

1 Introduction

In this paper, we address the problem of information retrieval and information extraction in subjective domains, with applications to visualization of opinions – specifically, thematic mapping of opinions. At present, there is a dearth of methods for integrating user profile data for social networks with blog posts, tweets, and other content from the associated social media. These limitations present an integrative challenge for human-computer interaction (HCI) and information retrieval (IR). Towards this end, the specific aims of the research proposed espoused in this position paper are as follows:

1. **Aim 1.** Extend known **algorithms for named entity recognition and relationship extraction**, to produce basic summaries of diseases and treatments mentioned in texts. The technical objective is to tag where basic entities and opinions are mentioned in freely available text (including both user posts and profiles), then map these tagged elements in space, time, and by topic, to acceptable levels of precision and recall.
2. **Aim 2.** Adapt basic known techniques to the domain of type 2 diabetes – specifically, extracting data from text discussions of diabetes that are archived from health blogs and forums using web crawlers. [1] This entails developing a means of handling entities and quantitative data that have not previously been extracted from text, such as information concerning insulin and oral antidiabetic drug dosage, HbA_{1c} levels, *etc.* Another functional requirement is some mechanism for entity reference resolution, *e.g.*, abbreviations and synonyms, for known terms. Finally, a **domainspecific ontology** of relevant symptoms, disease attributes, complications, and treatments is proposed. For type 2 diabetes, this includes topics frequently discussed in health blogs and forums: food groups, meal plans, nutritional constraints, and conditions such as obesity that are linked to diabetes. This shall facilitate information retrieval applications such as question answering about meal plans recommended by primary care physicians and specialists.
3. **Aim 3.** Develop methods for sentiment analysis and improve existing ones, to **summarize opinions and discover patterns**. The technical objective is to relate demographic data extracted from text and profiles to qualitative data – namely, the polarity of text at the document, sentence, or aspect level, aggregated across demographic categories such as

geographic region of residence. Objects of interest for sentiment analysis include prescribed therapies and specifically side effects, but can extend to disease aspects and complications.

The **overall goal** of this approach is to develop an integrative technology for summarizing online text about chronic diseases, capturing opinions from users' posts and demographic data from a combination of their posts and profiles, and finally using these to discover global patterns indicated by the set of all text documents. The **central hypothesis** of this work is that a combination of entity and relationship extraction, driven by a domain-specific ontology of terms, will result in more precise and accurate summarization of opinions. This will increase the usefulness of free-form text, written by users of social media, in understanding patterns that are reflected in the opinions and demographics of chronic disease patients.

2 Background

2.1 Information Extraction from Health Blogs

The chief potential impact of the research framework and test bed proposed in Section **Error! Reference source not found.** is to provide assistive technologies to public health analysts and health services analysts who are using blogs, microblogs (*e.g., Twitter*), and other social media to explore user opinions about chronic disease issues. As an example, in the application domain of type 2 diabetes, these include dietary treatments such as carbohydrate control, complications such as gastroparesis induced by diabetes that may pose digestive constraints, and recommendations of primary care physicians, therapists, endocrinologists, nutritionists, *etc*.

The availability of mailing lists, blogs, wikis, and other electronic media for content management and dissemination has resulted in rapid growth in the volume of online text data containing voluntarily expressed public opinions about health issues. While general-purpose metadata tools exist for annotating this text, the opinions themselves remain a largely unexplored source of information about how chronic diseases affect populations. Meanwhile, the task of relating content from these various self-publishing media to semi-structured profile data from their users has not yet been effectively automated.

We advocate development of application test beds and experimental systems aimed at improving techniques for information extraction, ontology development and mapping, and text mining to identify opinion patterns. The potential progress in these areas is due in part to the approach of combining information extraction to discover disease mentions with sentiment analysis to establish opinions, and in part to the application of this approach to a new source of data: free-form text describing user demographics, attributes of the chronic disease of interest and its related entities, and opinions and semi-structured profile data.

To help public health researchers tap into these freely available but unexplored sources of opinions, we propose to develop information extraction (IE) and summarization methods geared at health blog postings and similar text. Such postings contain not only opinions, and attribution information that can be used to link them to the users who expressed them, but also factual data about the posters and their opinions. This data can help place opinions in a comparative context [2] with population statistics, such as the reporting frequency of symptoms, side effects, and complications.

2.2 Ontology Development

The research approach centers around using information extraction to obtain structured data in the form of records about chronic disease references in text, which are then linked to users via relational data extracted from their profiles. [3] However, the body of relevant concepts in the healthcare domain and in the clinical domain theory of each chronic disease is much broader. Currently there exist preclinical (genomic and proteomic) and clinical translational ontologies [4] that contain information relevant to diabetes, but they do not provide the requisite concepts for mining free-form text written by lay users who are discussing diabetes online. We propose to develop an ontology for text mining in diabetes, and the mappings from extracted entities and relationships into this ontology.

2.3 Opinion Mining (Sentiment Analysis)

This aspect of the proposed work focuses on a basic research problem: sentiment analysis from text, also known as *opinion mining*, whose objective is to determine from analysis of a written document what the author's attitude towards an identifiable topic is. This attitude can be subjective or objective; it can be identified as an evaluation (positive or negative), a declaration of the author's emotional attitude, or a expression intended to evoke an emotional response in the reader. Subjects of interest include chronic diseases, their features or aspects including symptoms, complications, and treatments, and related health services.

2.4 Current State of the Field

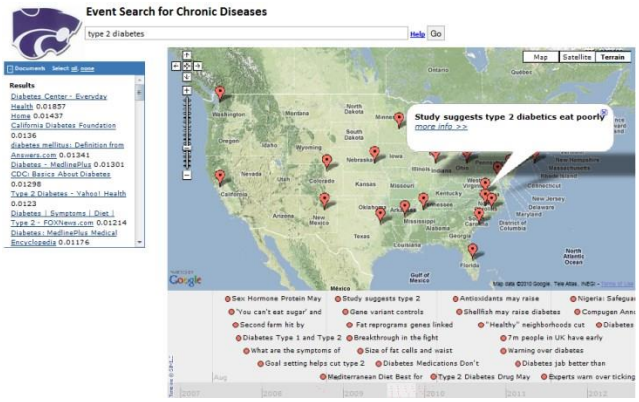

Figure 1. Prototype event search based on a previous IR system for veterinary epidemiology.

Figure 1 depicts a simple search interface for an existing IR system developed by the principal investigator's research group. This system was designed for event extraction in the domain of viral zoonoses, but uses general-purpose software for web crawling and ranking (the latter is developed using

Lucene Java). One marker is displayed on both the thematic map and the timeline for each returned page, but the only features extracted by this system are the disease name, formatted dates and times given in each article, and locations mentioned in the article.

The thematic map suggests several interactive functions related to opinion mining. One is content-based filtering of articles using the first type of thematic data, demographic and biostatistical attributes; another, collaborative filtering using the second type, polarity scores. Both of these use associations that can be learned from data: a user can search for articles by entering queries that express certain sentiments. In the first case, entities and attributes (*e.g.*, symptoms, complications, and treatments) mentioned in the query may match frequent patterns in the data; in the second, polarity scores themselves can be used to retrieve their "nearest neighbors in opinion space".

3 Example: Health Blogs

The primary value added in adapting the IR and IE workflows described above is an increased capability to explore patterns and trends expressed by an entire collection of health blog posts. As a running example, consider the public health analyst who is interested in charting trends in the use of fast-acting insulin by diabetics. Users often share information about the brands of insulin they use and post opinions about their effectiveness. The following is a post archived on *diabetesforums.com* which is marked up (with a color coding to distinguish entity types and relationship types):

> *Since I have found out in a previous thread I posted I can use most* ***pen needles*** *with the* ***Novolin 4 pen*** *I got(Haven't used it yet since I have only* ***Humalog*** *for rapids so far), I am here with another question.*
>
> *I have only used* ***Humalog*** *for a rapid... Does anyone have any* ***insight*** *as to* ***how it compares*** *to* ***Novolog****?*

In this post, the user is requesting information comparing two **drug products** (brands of fast-acting insulin, referring to specific **delivery mechanisms** (pen syringes), eliciting **opinions** from fellow users, and specifying a requested **comparison** between named products. Opinions voiced by respondents to this post then discuss how heat-tolerant each brand is, how quickly it acts, and other aspects we refer to as *facets*. [5] Achievement of our primary aims will allow analysts to chart reported biostatistics and opinions, not only about products but about trends, such as the number of units of postprandial insulin taken per gram of CHO.

4 Proposed Directions

4.1 Mining Social Media (Blogs, Lists, Wikis)

As mentioned above in Sections 1 and 2.1, our approach applies text mining to blogs and social media, a new source of information that is beginning to be studied for opinion and trending topic data, but has not been analyzed for disease-related information that can be related to these data. The novelty of our approach is that it extends named entity recognition and relationship extraction to the domain of understanding free-form text about aspects of chronic diseases (specifically, opinions about type 2 diabetes, its complications, dietary recommendations, and drug treatments). It further develops methods for mapping these new entities and relationships to the terms of an ontology for text mining, and finally leverages the text contained in many online sources to produce integrative summaries of disease mentions and associated opinions.

4.2 New Theory and Methodology

IR (Search Query-Driven) Workflow

In IR applications of automatic text summarization, a user enters a free-form search query and views returned hits that are summarized by topic and aspect – in this case, author opinion. These hits may be organized by space and time. For example, consider the case of a clinical health services analyst, public health analyst, doctor, patient, or other concerned individual who is interested in some aspect of a chronic disease. Such a user typically enters a query into a general-purpose search engine and is either directed to a domain-specialized web portal, also called a *vertical portal*, or browses through documents housed in one.

We seek to advance the state of the field by supporting structured queries, in which a user specifies fields and constraints in addition to traditional search keywords. This is achieved by combining quantitative text summarization (extraction of attribute values) with recognition of entities and relationships. The collection of documents may include some that are dynamically crawled from the web in response to the query. A mixture of labeled and unlabeled data is used to train a semi-supervised topic model. [6] The output consists of structured tuples that are ranked by relevance to the query, filtered to remove hits deemed insufficiently relevant, and finally visualized in a map or timeline view. This view allows the user to more freely explore information by performing interactive manipulations such as online analytical processing or editing the set of constraints.

IE and Summarization (Push) Workflow

IE applications of the proposed summarization technology can be viewed as a more passive variant of the IR application described above, from the user's point of view. [7] No initial query is supplied by the user, but there is an implicit domain of interest from which records should be displayed, corresponding to a combined set of search terms and relevance criteria. When a small set of search terms is known, the IE application can be formalized as a general case of the IR application where "every possible query" is enumerated, multiple crawls are conducted in advance, and the union of all resulting hits is ranked and filtered.

Improved Access through Structured Queries and Opinion Pattern Mining

This workflow is designed to provide analysts with better access to spatiotemporal data. First, it supports approximate range queries, such as: "return records of persons with

fasting blood glucose levels close to the nondiabetic range of < 126 mg/dL". Second, it uses measures of semantic relatedness or similarity, *e.g.*, "return posts about adverse effects of Metformin whose expressed sentiments are closest to those in this post". Third, it extracts information in Steps 1 – 3 that in Step 4 can be used to generate *thematic maps*, which portray specific aspects of a geographic region. In this research, the themes fall into two categories: the first, demographic attributes and biostatistics specified by the ontology – some diseaseindependent, and some disease-specific; the second, quantized measures of opinion polarity (*i.e.*, degree of positive or negative sentiment). The increased support for flexible queries and thematic map generation, compared to IR without relationship extraction and sentiment analysis, will help reveal patterns in the data through interactive investigation

5 Technical Focus Areas

5.1 Analytical Methods

The following generic methods are applied in order to meet the functional requirements presented in the preceding section. We refer to them as cross-cutting because they are used in service to all of the technical aims: entity and relationship extraction, ontology development, and sentiment analysis.

Focused Crawling

In previous work on IE, applied to news summarization in the domain of veterinary epidemiology, we used a combination of topical and focused crawling. *Topical crawling* prioritizes pages to be crawled based on userprovided terms (*i.e.*, topics) and seeds (*i.e.*, links to initial pages), while *focused crawling* uses both terms and pages labeled as positive or negative examples of relevant documents. Once tag-formatted web documents (HTML or XML) are crawled, text must be extracted from them.

The on-demand IR system described above functions by passing the user query to a built-in web crawler that fetches hits from a commercial search engine (in this case, *Yahoo*). The results are combined with previously crawled documents, if any, and ranked and indexed as a whole.

Information Extraction

The state of the field in IE for web articles describing disease consists of: payload extraction (of text from HTML), baseline named entity recognition (NER), and extraction of dates, times, and locations in order to localize putative events. In addition to general open natural language processing problems such as co-reference resolution (in particular, pronouns and other anaphora), word sense disambiguation, and canonicalization of dates, other IE problems that remain unsolved include: resolving alternative abbreviations and synonyms for diseases, disambiguation of place names, associating quantities of persons affected with diseases mentioned, and deduplication of reports. The foundation of our proposed work consists of tasks known to be feasible, but for which general-purpose solutions are still being manually adapted to new domains in current practice: automated named entity recognition and topic categorization. Typically, information extraction is restricted to named entities (Person, Organization, Location, and in our domain, Disease), but attributes such as "causative agent" are not always extracted. Neither are dates, times, quantities, and place names that support the extraction of full tuples of a relationship set. This open problem is of critical significance and is therefore the first of our specific aims.

Web 2.0

The term *Web 2.0* describes an eclectic set of technologies for online interoperability and collaboration. While it includes search, hyperlinking, collaborative authorship and tagging, web services, and syndication, our IE approach focuses on the **authorship**, **tagging**, and **syndication** aspects. Collaborative authorship and editing are mainstays of specialized wikis, but many forums also provide tools for collaboration, from discussion threading and editing history to user profiles, our main source of demographic information besides posts. We will crawl or aggregate profile data, which in some social network and blogging systems (*e.g.*, *LiveJournal*) is published as a publicly available feed. [8] Another source of relational data is the link structure expressed by collaborative tagging, especially annotation by other users *cf. Wikipedia*, social bookmarking *cf. Delicious*, social citation *cf. CiteULike*, and collaborative recommendation *cf. Digg*, *Reddit*, and *StumbleUpon*. We intend to make use of available content management functionality in health wikis and electronic groups. [9] Syndication provides a modern mechanism for refreshing content that is generally more efficient than periodic crawls. We will make use of these three categories of Web 2.0 features and other available content management functionality to assist in the extraction of relational tuples from free text writings online, and in their validation and ranking.

5.2 Visualization Methods

Maps and Timelines

Finally, the generation of views as shown in Figure 1 is a key application of our other primary aims: to build a domain ontology for text mining in diabetes blogs and develop automated mappings from entity recognition systems to this ontology; and to extract the objects and polarity of opinions.

Thematic maps, including opinion maps, help reveal **global patterns and trends** that may have been previously hidden. By visualizing the attributes and related entities of a disease and depicting their variation across space and time, they allow the user to interactively discover these trends. Most previous approaches to construction of thematic maps have been based on electronic medical records and reports compiled by medical providers or observers, such as individual incident reporters for *HealthMap*. [2] The value added by IE operations that automatically populate databases and thematic maps is that they can be applied to the large volume of text that is voluntarily submitted on a daily basis to venues listed at the beginning of this section.

References

[1] Jiang, J., & Zhai, C. Instance Weighting for Domain Adaptation in NLP. Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007), (pp. 264-271).

[2] Kim, H. D., & Zhai, C. (2009). Generating Comparative Summaries of Contradictory Opinions in Text. Proceedings of the 18th ACM International Conference on Information and Knowledge Management (CIKM 2009), (pp. 385-394).

[3] Jiang, J., & Zhai, C. (2006). Exploiting domain structure for named entity recognition. Proceedings of the Human Language Technology Conference/the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2006), (pp. 74-81)

[4] Craven, M., & Kumlien, J. (1999). Constructing biological knowledge bases by extracting information from text sources. Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology (pp. 77-86). Menlo Park, CA, USA: AAAI Press.

[5] Ling, X., Mei, Q., Zhai, C., & Schatz, B. R. (2008). Mining multi-faceted overviews of arbitrary topics in a text collection. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2008), (pp. 497-505).

[6] Yue, L., & Zhai, C. (2008). Opinion Integration Through Semi-supervised Topic Modeling. Proceedings of the 17th International World Wide Web Conference (WWW 2008).

[7] Yangarber, R., Steinberger, R., Best, C., von Etter, P., Fuart, F., & Horby, D. (2007). Combining Information Retrieval and Information Extraction for Medical Intelligence. NATO Advanced Study Institute on Mining Massive Data Sets for Security.

[8] Aljandal, W., Hsu, W. H., Bahirwani, V., & Caragea, D. (2009). Ontology-Aware Classification and Association Rule Mining for Interest and Link prediction in Social Networks. Proceedings of the AAAI 2009 Spring Symposium on the Social Semantic Web. Menlo Park, CA, USA: AAAI Press.

[9] Brownstein, J., & Feifeld, C. (2007). HealthMap – Global Disease Alert Mapping System. Retrieved January 25, 2010, from http://www.healthmap.org.