Journal of Biomedical Informatics 44 (2011) 536–544

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Leveraging health social networking communities in translational research

Yue W Webster [a,b,*], Ernst R Dow [a,b], Jacob Koehler [b], Ranga C Gudivada [b], Mathew J Palakal [a]

[a] School of Informatics, Indiana University Purdue University, IN, USA
[b] Discovery Informatics, Eli Lilly, IN, USA

ARTICLE INFO

Article history:
Received 10 March 2010
Available online 1 February 2011

Keywords:
Translational research
Health social networking
Semantic Web
Graph algorithm

ABSTRACT

Health social networking communities are emerging resources for translational research. We have designed and implemented a framework called HyGen, which combines Semantic Web technologies, graph algorithms and user profiling to discover and prioritize novel associations across disciplines. This manuscript focuses on the key strategies developed to overcome the challenges in handling patient-generated content in Health social networking communities. Heuristic and quantitative evaluations were carried out in colorectal cancer. The results demonstrate the potential of our approach to bridge silos and to identify hidden links among clinical observations, drugs, genes and diseases. In Amyotrophic Lateral Sclerosis case studies, HyGen has identified 15 of the 20 published disease genes. Additionally, HyGen has highlighted new candidates for future investigations, as well as a scientifically meaningful connection between riluzole and alcohol abuse.

1. Introduction

Health social networking communities (HSNC) are online communities where users search, self-track, share and discuss health-related information using Web2.0 technologies. Examples of popular HSNC include PatientsLikeMe.com (PLM), DailyStrength.org and MedHelp.org. Their primary users are patients with similar medical conditions. Content embedded in HSNC is adding a new category and dimension of information for translational research. For example, 5% of all Amyotrophic Lateral Sclerosis (ALS) patients in the US are registered members of PLM [1]. The data generated by those patients could arguably be the largest data set for ALS genotype–phenotype research. However, few studies have been devoted to explore the potentials and challenges in using HSNC as data sources in translational research.

The first challenge in handling HSNC content is the consumer-professional vocabulary gap. Vocabularies used by patients in HSNC are consumer English. On the other hand, most biomedical databases and tools are intended for professionals and use scientific vocabularies. For example, PLM allows patients to describe their conditions using folksonomy, a user-generated taxonomy. Less than half of those symptoms mapped to the concepts or synonyms in Unified Medical Language System (UMLS) [2].

Secondly, information in HSNC is organized and stratified by consumers using Web2.0 tools such as collaborative filtering, tagging, and voting. The data schema is derived bottom-up from the data, thus reflecting how patients understand and categorize biomedical knowledge. But when building traditional biological databases, we design the schema first and then load the data. Therefore, it is not a surprise that 62% of the symptoms submitted by PLM patients were not "Signs or Symptoms" in UMLS [2].

With over 20 large HSNC websites being launched in the last few years [3], there is an increasing need for novel tools and methods to address these challenges. We previously proposed a prototype for identifying hidden associations related to colorectal cancer using information extracted from traditional biomedical databases [4]. Based on the initial prototype, we have designed and implemented novel strategies to overcome the challenges in HSNC content. This manuscript presents the completed framework, named HyGen, focusing on how the community-level data in PLM is processed and utilized. In addition, it describes our quantitative evaluations, proposes an optimization method, and discusses the preliminary results.

2. Related work

Graph analysis has drawn much interest among bioinformatics researchers due to the rapid growth of publicly available high throughput data [5–12]. Such data have provided linkages among

Abbreviations: HSNC, Health social networking communities; ALS, Amyotrophic Lateral Sclerosis; MFR terms, most frequently reported terms; PLM, PatientsLikeMe.com.

* Corresponding author. Address: DC 1930, Lilly Corporate Center, Indianapolis, IN 46285, USA. Fax: +1 317 276 6545.

E-mail addresses: yuwang@umail.iu.edu (Y.W Webster), dow@lilly.com (E.R Dow), kohlerja@lilly.com (J. Koehler), gudivadara@lilly.com (R.C Gudivada), mpalakal@iupui.edu (M.J Palakal).

1532-0464/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2011.01.010

chemical, biological, and clinical entities, which can be modelled as nodes and their relationships as edges (links). The graphs can then be analyzed using conventional graph analysis technique or extension of it [11–15].

In this case, we are especially interested in applying graph algorithms to rank search results. Common approaches [16–19] include concept structure analysis, PageRank, and Hyperlink-Induced Topic Search (HITS). PageRank with Priors proposed by White and Smyth [20] simulates the steps of a Web surfer, who starts from any of the root nodes on the Internet and follows a random link at each step with β as the probability of returning to the root nodes. A score is computed for each node on the Internet to reflect its probability of being reached by the surfer. This score is used to measure the relative "closeness" of a node to the root nodes. K-Step Markov method simulates a similar Web surfing scenario as in PageRank, except that the surfer returns to the root nodes after K steps and restarts the process. K-Step Markov algorithm estimates the relative probability that a surfer will spend time at a node given that the surfer starts in a set of root nodes and stops after K steps. HITS with Priors proposed by Kleinberg measures two properties of a node: (1) authority score estimates the importance of the node itself; and (2) hub score measures the importance of other nodes linked to the current node [21]. Therefore, HITS with Priors not only considers the number of links to and from a node but also its neighbours'.

Gudivada et al. have proposed a modified algorithm to rank genes [22]. In traditional WWW ranking analyses, all links are considered equally significant. But in the context of biological networks, the importance of a link also depends on the nodes connected with it. Using gene and pathway association as an example, Gudivada explained that a gene participates in multiple pathways is more important than a pathway that has multiple genes since most pathways will include multiple genes. To model this nature of biological networks, each link is assigned a subjectivity weight and an objectivity weight. Link such as '*Gene-HasAssociated-Pathway*' is assigned a higher subjectivity weight (for gene) and lower objectivity weight (for pathway). The only constraint is that for each link the sum of subjectivity and objectivity weights must be equal to 1.

Although concepts and technologies supporting semantic ranking have been studied by many researchers mentioned above, fewer reports have been published on applying user profiling and sub-graphing technologies to rank multi-level and cross-disciplinary biomedical data based on graph attributes. This approach integrates many types of nodes to discover associations among different types of biomedical entities and to deliver the results based on each user's interest.

3. Methods

HyGen combines Semantic Web (SW), graph algorithm and user profiling to discover novel associations. The discovery process has two main steps: (1) constructing a full semantic graph using associations extracted from heterogeneous sources; (2) subtracting a sub graph and ranking the associations based on the criteria defined by the user.

3.1. Constructing a weighted graph

Using the method proposed in [23], associations were extracted from well-known genomic, pharmacological and proteomic databases. A summary of the compiled associations is displayed in Table 1. Those associations were converted to nodes and edges in the full graph, where nodes represent life science entities, such as genes, diseases, or compounds; and edges represent the relationships between entities. Numerical weights (from 0 to 1) were assigned to the edges based on the confidence scores of the data sources. Users can adjust the confidence score of a source according to their own experience and needs in the user profiles.

Table 1
Associations in the full graph.

Associations	Count	Source database
Gene and clinical features	150,292	OMIM (www.ncbi.nlm.nih.gov) GAD (http://geneticassociationdb.nih.gov) PharmGKB (www.pharmgkb.org)
Gene and gene	310,842	BioGrid (thebiogrid.org) BIND (bond.unleashedinformatics.com) MINT (mint.bio.uniroma2.it) IntAct (www.ebi.ac.uk/intact) Reactome (reactome.org)
Gene and pathway	91,771	KEGG (www.genome.jp/kegg) Reactome (www.reactome.org) WikiPathways (www.wikipathways.org) Panther (www.pantherdb.org) PID (pid.nci.nih.gov) GeneGo (www.genego.com)
Drug and gene	6552	DrugBank (www.drugbank.ca) PharmGKB (www.pharmgkb.org)
Drug and clinical features	6742	DrugBank (www.drugbank.ca) PharmGKB (www.pharmgkb.org)

In the full semantic graph, the Uniform Resource Identifier (URI) of an entity is derived from NIH authoritative identifiers, such as EntrezGene ID or UMLS CUI. Thus, entities sharing the same URI are merged into one node regardless of their sources, and two nodes from different disciplines are associated if they both connect to the same node. It is worth mentioning that even though we used Semantic Web technologies here, the full graph can be constructed by any other technologies as long as they allow HyGen to merge and connect entities from diverse sources.

3.2. Converting patient-reported terms to nodes

In PLM, individual-level data is aggregated and reflected in the community reports. We extracted the most frequently reported (MFR) symptoms from ALS community report provided by PLM [24]. We normalized the symptom terms against UMLS [25] using MetaMap [26], followed by manual inspection. The top ten MFR symptoms and the matching UMLS concepts are displayed in Table 2. Similarly, we extracted MFR prescription drug names from the ALS community report [27] and normalized them against compounds or synonyms in CHEMLIST, a dictionary for identifying chemical information in the literature [28].

Thus each patient-reported term was mapped to the node that represents the same clinical concept or chemical substance in the full graph. Mapping of instance-level data is HyGen's key strategy to bridge the consumer-professional vocabulary gap. Ontologies such as UMLS, CHEMLIST and their companion linguistic tools, in combination with SW, have made it possible to aggregate HSNC content with data from research-oriented resources.

The other challenge mentioned previously is that more than half of the symptoms submitted by PLM patients were not "Signs or Symptoms" in UMLS. We circumvented this problem by defining one general type called "clinical-feature" for concepts belonging to multiple UMLS semantic types. The relationships between clinical-features and other types of nodes were loosely defined (e.g. "related_to_gene" and "related_to_drug"). Obviously, the penalty of this approach is a higher false positive rate. To compensate, we implemented a pseudo relevance feedback strategy to reduce the irrelevant connections.