A Semantic Approach to involve Twitter in LBD Efforts

Sanmitra Bhattacharya and Padmini Srinivasan
Department of Computer Science
The University of Iowa
Iowa City, Iowa 52242
Email: {sanmitra-bhattacharya, padmini-srinivasan}@uiowa.edu

***Abstract*—Literature-based Discovery (LBD) refers to the task of finding hidden, unknown or neglected relationships that may be uncovered using biomedical text. While traditional LBD primarily focuses on MEDLINE records for unearthing such relationships, recent studies have shown the applicability of contemporary textual resources such as electronic medical records or online medical message boards for similar purposes. In this paper we highlight yet another source for LBD, i.e., Twitter data. We focus on the use of Twitter as a new resource for finding hypotheses — both novel and slightly studied. Using a set of drug and disease names as starting points we retrieve thousands of Twitter messages which are then processed for semantic information to mine several hundred biomedical relationships which we call probes. Manual inspection of a handful of these probes reveals instances where tweets strongly support a hypothesis for which no evidence can be found in PubMed. In other cases, we find very few related PubMed records supporting/rejecting such Twitter-mined probes. Overall, we show the importance and usefulness of Twitter for LBD efforts.**

***Index Terms*—Literature-based discovery, semantic information, Twitter, public health informatics.**

I. Introduction

The area of Literature-based discovery (or LBD) has operated largely off bibliographic data as in PubMed or full-text collections as for example PubMed Central. Allied discovery efforts using resources such as electronic medical records [1], annotation data [2] and online medical message boards [3] have also been explored for LBD. Our goal here is to investigate a source that is, to the best of our knowledge, unused for LBD. We investigate methods for using social media, and in particular, Twitter data as a source for ideas that may feed into the LBD process. Our motivation is that many diverse conversations on social media often include observations posted by patients, their family members and friends. Considered in aggregate these observations could have the potential to be the basis for new ideas and hypotheses.

The approach we propose involves two phases (see Figure 1). In the first phase we mine Twitter looking for specific types of statements, namely, those about treatments and causes of diseases. In the second phase we determine if the statement has already been discussed in the scientific literature, i.e., in PubMed. Thus we propose using Twitter, which holds public discussions of health and illnesses, as an alternative source for new ideas. As in traditional LBD work, we continue to use PubMed to verify or validate the novelty of the ideas.

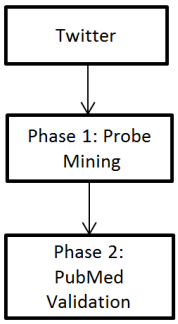

Fig. 1. Two phases of LBD from Twitter

Our aim in this work is to present a method that has the *potential* to enrich LBD efforts and our aim is also to identify the challenges that remain in this direction. We also discuss several findings (in the form of 'case studies') from our 2 phase process.

The Twitter mining methodology we propose involves semantic processing of Twitter posts or tweets. This involves identifying concepts belonging to specific semantic categories, instances of relationships of particular semantic types and ensuring that the tweet indeed conveys the right meaning. We find that semantics is a necessary and also a highly challenging component of our method. Hurdles in language, spam, noise, high levels of ambiguity need to be handled. We also note that the semantic processing used in this paper comprises of both automated processing using semantic tools or methods and manual post-processing based on semantic information conveyed in the tweets.

II. Related Research

Literature-based discovery has a long-standing history and has been studied extensively in the past few decades. Starting with Swanson's seminal paper on "undiscovered public knowledge" [4] and the discovery of the hidden association of fish oil and Raynaud's disease [5], LBD has seen extensive growth across multiple dimensions. Several other hypothesis were discovered over the years following the basic paradigm proposed by Swanson, including the relationships between migraine and magnesium [6], Somatomedin C and arginine[7], etc.

Also more sophisticated methods, like the use of MeSH terms [2][8] and semantic technologies [9][10][11] rather than the simple term co-occurrence-based method used in Swanson and Smalheiser's initial approach [12], have been used to facilitate the LBD process.

More recently, LBD has seen an increasing use of contemporary resources such as internet message boards, electronic media and electronic medical records. For example, mining a corpus of breast cancer message boards have been shown to be effective in finding novel adverse drug effects that were otherwise not found in package labels [3]. Similarly, the the association of Vioxx with myocardial infarction has been shown to be identifiable from a corpus of Google News long before the findings were widespread [13]. The use of social media, particularly Twitter, has been limited in mining hypothesis for LBD purposes. In this paper we propose a novel method for involving Twitter in LBD using semantic technologies followed by PubMed validation.

III. Semantic Mining of Twitter for Drug and Disease Information

Our focus is on mining Twitter discussions about drugs and diseases. However, our method is general and may be applied to other discussions such as on organisms or surgical interventions. We narrow our interest further to focus on a) the effects of drugs, which includes both positive and side or negative effects and b) the causes and treatments of diseases. In other words we develop a semantic method to identify discussions on topics of interest and mine them for particular semantic relations (described next).

Given a set of tweets that discuss a particular topic (say a drug X) we are interested in mining binary semantic relations connecting X with key concepts (in this example possibly disease concepts). This goal of extraction is not, in itself, a novel idea. Established systems such as SemRep ([14], [15]) are there to extract relations from texts and have been applied fairly extensively to MEDLINE ([16], [17]). The novel aspect in our work is in the extraction of biomedical relations from Twitter data. Twitter, notorious for its casual language, offers extreme challenges in levels of noise, ambiguity and stylistic variants. Thus the accurate extraction of binary relationships from Twitter that fit particular semantic constraints is not to be taken for granted. A further novel aspect is that we connect this extraction step with LBD (as shown in Figure 1). In other words, we include a validation step (PubMed search) to see if the extracted relationship already appears in MEDLINE. Note that in terms of nomenclature we refer to the extracted binary relationships as probes.

A. Phase 1 Details: Mining Probes from Twitter

Our phase 1 goal is to mine a set of probes (relationships) for a given set of drugs and diseases (see Figure 2). We start with a list of drugs and diseases (Table I). For drugs we take the top 10 most Direct-to-Consumer (DTC) advertised drugs[1] for heart diseases, neuropathic pain, etc. We also selected 10 over-the-counter (OTC) or generic drugs[2] that are less advertised but frequently used for common problems like fever, pain, heartburn, etc. For diseases we selected chronic[3] and infectious diseases[4] from the World health Organization's (WHO) fact sheets for such diseases.

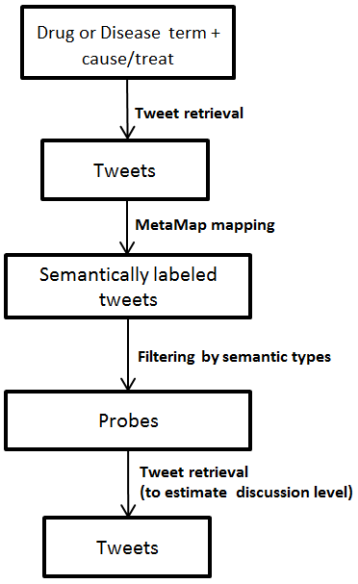

Fig. 2. Flowchart of Probe Mining (Phase 1) from Twitter

TABLE I
Drugs and Diseases Explored

Top 10 DTC Drugs	Lipitor, Cialis, Advair, Abilify, Cymbalta Symbicort, Pristiq, Plavix, Chantix, Lyrica
OTC Drugs	Aspirin, Advil, Prilosec, Centrum, Robitussin Tylenol, Nyquil, Dramamine, Zantac, Benadryl
Chronic Diseases	Diabetes, Asthma, Arthritis, Schizophrenia, Cardiac failure Glaucoma, Haemophilia, Hypertension, Multiple sclerosis, Parkinson's disease, Osteoporosis, Psoriasis, Obesity, Epilepsy
Infectious Diseases	HIV/AIDS, Dengue, Malaria, Anthrax, Cholera Bubonic plague, Influenza, Typhoid, Smallpox, Pneumonia Tuberculosis, Yellow fever, Bird flu, Ebola, Leprosy, Hepatitis

1) Tweet retrieval for initial set of probes: As a first step we retrieve relevant tweets discussing the relationships of interest for our particular drugs and diseases. For this we combine each drug and disease term (in turn) from Table I with the relationship terms *cause* and *treat* (and their plural variants) and use these combinations to search on Twitter. Thus for the 50 terms in Table I, we conducted a total of 200 searches[5] (for details on the search strategy see [18]). We remove URLs, user mentions and re-tweet mentions and take the unique instances of the remaining tweets. For our sets of drugs and diseases this resulted in 942 and 3722 unique tweets respectively. Each of the 942 'drug' tweets and the 3722 'disease' tweets has both a focus drug (or disease) name and a relationship term.

[1] http://gaia.adage.com/images/bin/pdf/WPpharmmarketing_revise.pdf

[2] http://www.uihealthcare.com/pharmacy/OTCmedications.html

[3] http://www.who.int/topics/chronic_diseases/factsheets/en/

[4] http://www.who.int/topics/infectious_diseases/factsheets/en

[5] This was done on February 24, 2012 using the Twitter Search API

2) Semantic processing of tweets & filtering: Next these tweets were processed using the National Library of Medicine's MetaMap program [19] (with word-sense disambiguation option) to identify biomedical concepts and their semantic types. We then select those tweets that had the focus drug (or disease) along with concepts of a pre-specified set of semantic types. Note that these tweets also have the relationship term: cause or treat as per our retrieval strategy and are thus expected to represent the types of Twitter discussions of interest to us.

To elaborate, for 'disease' tweets we kept tweets that also had at least one concept from [Organic Chemical], [Pharmacologic Substance], [Amino Acid, Peptide, or Protein], [Food], [Manufactured Object], [Mammal], etc. When analyzing 'drug' tweets we kept those that also had at least one concept from [Sign or Symptom], [Disease or Syndrome], [Mental or Behavioral Dysfunction], etc. We then extracted pairs of concepts, one being the focus drug (or disease) and the second being the concept falling into these pre-specified semantic types. We refer to these pairs (along with the particular relationship term) as probes. Using this strategy we identified 361 and 978 unique probes for the drug and disease sets respectively.

Manual inspection of these probes revealed certain anomalies. For example, for the tweet "The girl in my class is giving a speech and said weed causes schizophrenia", MetaMap identifies the verb 'said' as 'said (Simian Acquired Immunodeficiency Syndrome) [Disease or Syndrome]'. Other examples of frequently appearing but incorrectly mapped terms are 'I' (identified as 'I NOS (Blood group antibody I) [Amino Acid, Peptide, or Protein], [Immunologic Factor]'), dnt (abbreviated don't) (identified as 'DNT (Dysembryoplastic neuroepithelial tumor) [Neoplastic Process]'), etc. On the other hand, we also identified some very interesting probes such as 'armadillos[6] cause leprosy' because of the use of wider range of semantic types. After manually filtering out such incorrectly mapped terms we had 209 and 556 probes from the drugs and diseases datasets respectively. Table II summarizes these dataset characteristics. We note that disease-related search terms retrieve more tweets than drug-related terms. Consequently, the number of probes mined from disease-related tweets is greater than that of drug-related tweets dataset.

3) Tweet retrieval to estimate level of discussion: Note that our initial Twitter search had been limited to the relationship terms *cause, causes, treat* and *treats*. That search was deliberately directed towards high precision as the aim was to extract probes about causes and treatments. But to estimate extent of discussion we need a retrieval strategy that emphasizes recall with minimal sacrifice of precision. It is easy to observe that Twitter users offer many variant expressions of a single idea. To illustrate, the 'treats' relationship may be expressed with phrases such as 'helped me recover from' or 'made me feel better'. Thus we adopt a different recall-emphasizing strategy for retrieving relevant tweets as described in [18][7]. 117 probes of the drugs dataset and 324 probes of the diseases dataset retrieved more than 10 tweets (last row of the table). Note that we consider original tweets as well as re-tweets of the original tweet in these counts of discussions on a particular probe.

[6]Armadillos (*Armadillo officinalis*) are mammals primarily found in Central and South America.

TABLE II
DRUGS AND DISEASES DATASETS

Dataset: *Drugs*	
Number of tweets retrieved for Disease set	1226
Number of unique tweets	942
Number of probes (before filtering)	361
Number of probes (after filtering)	209
Number of probes retrieving $>$ 10 tweets	117
Dataset: *Diseases*	
Number of tweets retrieved for Drugs set	8679
Number of unique tweets	3722
Number of probes (before filtering)	978
Number of probes (after filtering)	556
Number of probes retrieving $>$ 10 tweets	324

TABLE III
SAMPLE SET OF MINED PROBES

Sample Probe Statements	
Advil treats hangover	Ginkgo treats diabetes
Advil causes stomach bleeding	Marijuana causes schizophrenia
Armadillos causes leprosy	Methotrexate treats cancer
Video causes seizure	Bergamot treats psoriasis
Benadryl causes itching	Nigella sativa treats diabetes
Bilberry treats diabetes	Nyquil causes coma
Neem treats psoriasis	Weed treats depression
Coffee causes diabetes	Viagra causes hearing loss

A sample set of mined probes is shown in Table III. There were several probes involving alternative medicine (herbal therapy, homeopathy, etc.) and dietary substances having causal or curative relationships with diseases. Not surprisingly, we find a large number of probes relate to recent studies with animal models that might have generated a buzz. Naturally probes mined depend upon current events and developments. This is because social media often correlates to current events in news media or even pop-culture. As an example several tweets yielding the probe "video causes seizure" refer to a popular music video that might cause epileptic seizure and contains a related disclaimer.

B. Phase 2 Details: Validation by PubMed Search

Validation of probes is challenging. Since our goal is hypothesis discovery there are two criteria to satisfy, namely novelty and rationale. Novelty refers to whether the idea underlying the probe has already been explored in the scientific research. One may of course adopt a broader perspective and assess if the idea has been even discussed or utilized outside of the scientific arena. The second criteria of at least equal importance is that of rationale or reasonableness. To what extent is the idea, unexamined as it may be, supported

[7]This Twitter search was performed on February 28, 2012; retrieved tweets dated back to the previous 7 days as per the API. We collected 88,048 tweets for the 765 filtered probes (Table II).

by the scientific literature? Some arguments on these issues, especially on novelty, have been presented most energetically by [20][21][22]. In the case of probes derived from social media, the criteria of reasonableness perhaps is even more crucial than with probes mined from peer reviewed collections or clinical records. Overall, the role of novelty, support, reasonableness, extent etc. in hypothesis discovery are intricate not just to measure but even to define; perhaps there are even elements of subjectiveness determined by the inclinations of particular scientific sub-communities. We take an initial validation step here by exploring the presence of these probes in MEDLINE through a PubMed search. Note that the probe's absence (or low presence) in MEDLINE does not necessarily indicate a reasonable or interesting hypothesis. But yet it is a start towards hypothesis discovery. One decision we can be confident about is that if a probe has an 'appreciable' MEDLINE footprint then one must remove it from further consideration.

We conduct our validation PubMed search using the following strategy. First we search the PubMed title/abstract fields using the concept terms limited to publications dated prior to Feb 28, 2012. For example the PubMed search for a probe such as 'aspirin treats poor leg circulation' is **(aspirin[Title/Abstract] AND poor leg circulation[Title/Abstract]) AND ("1800"[Date - Publication] : "2012/02/28"[Date - Publication])**. If we find multiple hits with this strategy then we add the relationship term 'treats[Title/Abstract]' to the search query. On the other hand, if we do not find any hits for the previous search strategy, we relax the search query to a simple keyword search using the concepts which helps us identify possible synonyms for the concepts. These are then replaced in the original query and search is again executed. It is important to note here that consumer vocabulary (as mined in probes) is significantly different from standardized or scientific vocabulary [23]. While the PubMed search algorithm implicitly does query expansion to include standardized or scientific terms for common terms, it is not comprehensive. The following search results are limited to the use of common terms used in the probes and alternative scientific terms suggested by PubMed.

IV. Analysis of Select Probes

We manually analyzed a sample of two groups of probes (relationships) based on the number of PubMed documents retrieved. In the first category, *Probes with Sparse PubMed Support*, we have probes that retrieved low to moderate number of PubMed records. In the second category, *Probes with No Explicit PubMed Support*, we have probes for which we could not find any PubMed records. When we did find records, we also took a look at them to see if the probe was being discussed or if it was a false positive retrieval.

A. Probes with Sparse PubMed Support

1) Curcumin treats multiple sclerosis (MS): Sample Tweet: "Curcumin has bright prospects for the treatment of multiple sclerosis [URL]"

While we find only few tweets retrieved for this probe, 10 articles on this probe are retrieved using PubMed search. Adding the relationship term 'treat' however results in only one article. A manual analysis of the abstracts of these articles reveal various indications of its benefits – from discussion of its efficacy in animal models to its anti-inflammatory properties in specific scenarios – without any concrete evidence of its use in curing MS in humans.

2) Cilantro treats diabetes: Sample Tweet: "Apparently cilantro is used to treat diabetes...well I hope to god I don't get it cause I can't stand cilantro."

Only the above tweet was retrieved for this probe. While a PubMed search of the exact concept term pairs did not return any results, replacing cilantro with its scientific name coriandrum resulted in 10 hits. These articles covered various topics including other traditional plant treatments for diabetes to its efficacy in animal models.

3) Coconut oil treats psoriasis: Sample Tweet: "RT @Psoriasisclub: Coconut Oil: a fantastic natural moisturiser for any dry skin and especially helpful for psoriasis. [URL]"

While 13 tweets referred to this probe, only 2 studies could be found using PubMed search. One of the studies found no significant benefits in using coconut oil for psoriasis clearance, while the other discusses the process of a drug preparation ("77 oil") which uses coconut oil as a base and used in the treatment of psoriasis.

4) Cialis causes hearing loss: Sample Tweet: "RT @Iamsuperbrad: One side effect of Cialis can be hearing loss. [expletive satire] It's every man's dream in pill form."

38 tweets were retrieved supporting this probe. A PubMed search on this probe (using generic name Tadalafil) resulted in 2 retrieved records both indicating hearing loss due to various Phosphodiesterase type 5 inhibitors.

5) Cialis treats high blood pressure: Sample Tweet: "cialis treat high blood pressure [URL]"

A large number of tweets (97) supporting this particular probe was mined from Twitter. Using our strict PubMed search strategy we did not find any evidence of this association. However using the relaxed search strategy we found several instances where Cialis (generic name Tadalafil) is used as a treatment for pulmonary arterial hypertension.

6) Krill oil treats Rheumatoid Arthritis (RA): Sample Tweet: "krill Oil Supplements Can Treat the Symptoms of Rheumatoid Arthritis [URL]"

We found around 10 tweets discussing this probe. A PubMed search of the probe returns two records. One of these records, an older study found "Neptune Krill Oil (NKO)" to be beneficial for RA, while a more recent study from 2010 demonstrates its efficacy in animal models.

7) Bergamot treats psoriasis: Sample Tweet: "Fischer-Rizzi sugests blending bergamot with rock rose and everlasting to treat eczema and psoriasis. #aromatherapy #skincare"

This probe results in only 3 PubMed hits. A manual inspection of the PubMed records reveal direct or indirect relationship between bergamot (specifically bergamot oil) and

psoriasis. A search of bergamot and eczema using the strict or relaxed PubMed search shows no results.

8) Cialis causes muscle pain: Sample Tweet: "cialis side effects muscle pain [URL]"

16 tweets were retrieved for this probe. However, a PubMed search with the generic name Tadalafil resulted in only 2 hits, both related to adverse effect for this drug.

9) Benadryl causes hallucinations: Sample Tweet: "OMG!!!! benadryl causes hallucination! #hallucinate"

This probe mined from Twitter fetches only 2 PubMed records when searched as-is. However, using 'diphenhydramine', the generic name for Benadryl, we get 19 hits. Appending the relationship term 'cause' to the search results in only 5 hits.

B. Probes with No Explicit PubMed Support

1) Lavender oil treats acne/psoriasis: Sample Tweet: "#Natural #Health: Lavender oil has been used for centuries to treat acne, wrinkles, psoriasis + skin irritants [URL] #beauty"

While quite a few tweets (8) were found relating lavender oil to treatment for acne and psoriasis, we did not find any PubMed records supporting such claims. However, 'wrinkles', which is also mentioned in the same tweet, retrieves one PubMed record when associated with Lavender oil. Manual inspection of the article reveals the use of lavender oil aroma for easing anxiety of patients undergoing BOTOX treatment for wrinkled skin.

2) Triphala treats obesity: Sample Tweet: "Triphala treats obesity miraculously. As triphala regularizes the functioning of our digestive system, it directly reduces body fat."

While a single tweet referred to this particular probe, PubMed search of triphala (an Ayurvedic medicine comprised of three myrobalans) and its treatment potential for obesity did not return any results.

3) Clove oil treats colds/bronchitis/asthma/tuberculosis: Sample Tweet: "Clove leaf oil is also clearing nasal passage & treat colds, bronchitis, asthma, and tuberculosis."

While we find evidence of this probe in Twitter discussions, PubMed searches of the association of clove oil with any of the diseases or symptoms do not return any result.

4) Neem treats psoriasis: Sample Tweet: "Using Neem to Treat Psoriasis — 21st Century Apothecary [URL]".

We did not find any documents relating neem *(Azadirachta indica)* with psoriasis.

5) Lyrica causes hair loss: Sample Tweet: "@CraigHeff Lyrica, Topamax, Lamictal are all used for neuropathic pain releif. Side effects are, suicidal thoughts, memory and hair loss."

While we find sparse evidences of association of Lyrica (generic Pregabalin) with suicidal thoughts in PubMed (13 tweets), there is no evidence of the adverse effect of hair-loss in association with Lyrica in PubMed search (2 tweets).

6) Cialis causes heartburn: Sample Tweet: "why does cialis cause heartburn [URL]".

41 tweets reporting this side-effect were found in our dataset. However a PubMed search of this probe using both the brand name and the generic name of the drug returned no results.

7) Ginkgo treats bronchitis: Sample Tweet: "[URL] Ginkgo leaves and seeds are utilized to treat asthma, bronchitis, allergies, cardiac arrhythmia and to improve memory".

A PubMed search of the various treatment related probes of Ginkgo, namely for treatment of asthma or allergy or cardiac arrhythmia , return at least a few articles. However no article could be found on the efficacy of Ginkgo for bronchitis.

8) Rosehip oil treats acne/eczema/psoriasis: Sample Tweet: "Rosehip oil used to treat stretch marks, burns, sunburn surgery scars, acne, eczema, psoriasis [URL]".

No evidence of associations between rosehip oil and any of the skin conditions listed was found in PubMed.

In summary, several probes or relationships were mined from Twitter that are either not present or sparsely present in PubMed. (Note this statement is made within the constraints of our search strategy). At first glance these probes representing proto-ideas have some potential towards developing new hypotheses for scientific research. However, these need further validation especially regarding reasonableness or rationale and this validation may involved downstream text mining processes. Social media being the source underlines this need. In fact, a natural strategy, which we propose for the future, is to put the mined probes through a closed discovery process [5] to extract any underlying rationale (for instance between Triphala and obesity). Overall we show that, our method exploiting the semantics of concepts, is capable of mining specific types of relationships from Twitter discussions that could feed into a a more general LBD process.

V. Conclusion

In this paper we have shown the use of Twitter data for finding various known and unknown biomedical hypothesis. Starting from a set of select drug and disease names we mined various relationships or probes from Twitter. A few of these were manually checked on PubMed for supporting evidences. While for some of these probes (e.g. Coconut oil treats psoriasis) we found explicit published evidence, a considerable number of probes (e.g. Neem treats psoriasis) remain unsupported or lacked explicit evidence on PubMed. Our goal in this paper was not to verify the scientific validity of such hypothesis but to simply show the use of contemporary and ever-growing channel of information propagation, that is social media, in LBD. Our current work is limited by the semi-automated semantic type-based processing of tweets to find meaningful probes. As per reviewer suggestion, we tested the use of SemRep [17] for mining probes on a small set of tweets. Using the 17 'Sample Tweets' presented in Section IV, we were able to mine only 8 probes using SemRep as opposed to 17 using our proposed approach. In future work we would like to implement more sophisticated techniques that can sift through huge amounts of Twitter data to find such probes more efficiently. In this paper, although we identified the probes generating more discussion ($>$10 tweets) than others (Table II), we have not studied the discussion levels of probes

w.r.t. specific semantic types. In future work we would like to explore which topics generate more discussion on Twitter.

REFERENCES

[1] F. S. Roque, P. B. Jensen, H. Schmock, M. Dalgaard, M. Andreatta, T. Hansen, K. Soeby, S. Bredkjaer, A. Juul, T. Werge, L. J. Jensen, and S. Brunak, "Using electronic patient records to discover disease correlations and stratify patient cohorts," *PLoS Comput. Biol.*, vol. 7, no. 8, p. e1002141, Aug 2011.

[2] P. Srinivasan, "Text mining: generating hypotheses from MEDLINE," *J. Am. Soc. Inf. Sci. Technol.*, vol. 55, pp. 396–413, March 2004.

[3] A. Benton, L. Ungar, S. Hill, S. Hennessy, J. Mao, A. Chung, C. E. Leonard, and J. H. Holmes, "Identifying potential adverse effects using the web: A new approach to medical hypothesis generation," *J. of Biomedical Informatics*, vol. 44, no. 6, pp. 989–996, Dec. 2011.

[4] D. R. Swanson, "Undiscovered public knowledge," *The Library Quarterly*, vol. 56, no. 2, pp. 103–118, 1986.

[5] D. R. Swanson, "Fish oil, Raynaud's syndrome, and undiscovered public knowledge," *Perspect. Biol. Med.*, vol. 30, no. 1, pp. 7–18, 1986.

[6] D. R. Swanson, "Migraine and magnesium: eleven neglected connections," *Perspect. Biol. Med.*, vol. 31, no. 4, pp. 526–557, 1988.

[7] D. R. Swanson, "Somatomedin C and arginine: implicit connections between mutually isolated literatures," *Perspect. Biol. Med.*, vol. 33, no. 2, pp. 157–186, 1990.

[8] M. Yetisgen-Yildiz and W. Pratt, "Using statistical and knowledge-based approaches for literature-based discovery," *J Biomed Inform*, vol. 39, no. 6, pp. 600–611, Dec 2006.

[9] D. Hristovski, B. Peterlin, J. A. Mitchell, and S. M. Humphrey, "Using literature-based discovery to identify disease candidate genes," *Int J Med Inform*, vol. 74, no. 2-4, pp. 289–298, Mar 2005.

[10] X. Hu, X. Zhang, I. Yoo, X. Wang, and J. Feng, "Mining hidden connections among biomedical concepts from disjoint biomedical literature sets through semantic-based association rule," *Int. J. Intell. Syst.*, vol. 25, no. 2, pp. 207–223, Feb. 2010. [Online]. Available: http://dx.doi.org/10.1002/int.v25:2

[11] X. Hu, "Mining novel connections from large online digital library using biomedical ontologies," *Library Management*, vol. 26, no. 4/5, p. 261, 2005.

[12] D. R. Swanson and N. R. Smalheiser, "An interactive system for finding complementary literatures: a stimulus to scientific discovery," *Artif. Intell.*, vol. 91, no. 2, pp. 183–203, Apr. 1997.

[13] D. MacLean and M. I. Seltzer, "Mining the web for medical hypotheses - a proof-of-concept system," in *HEALTHINF*, 2011, pp. 303–308.

[14] T. C. Rindflesch and A. R. Aronson, "Semantic processing in information retrieval," *Proc Annu Symp Comput Appl Med Care*, pp. 611–615, 1993.

[15] T. C. Rindflesch, J. V. Rajan, and L. Hunter, "Extracting molecular binding relationships from biomedical text," in *Proceedings of the sixth conference on Applied natural language processing*, ser. ANLC '00. Stroudsburg, PA, USA: Association for Computational Linguistics, 2000, pp. 188–195. [Online]. Available: http://dx.doi.org/10.3115/974147.974173

[16] D. Hristovski, C. Friedman, T. C. Rindflesch, and B. Peterlin, "Exploiting semantic relations for literature-based discovery," *AMIA Annu Symp Proc*, pp. 349–353, 2006.

[17] T. C. Rindflesch, P. D, and A. R. Aronson, "Semantic processing for enhanced access to biomedical knowledge," in *In Real World Semantic Web Applications*. IOS Press, 2002, pp. 157–172.

[18] S. Bhattacharya, H. Tran, P. Srinivasan, and J. Suls, "Belief Surveillance with Twitter," in *Proceedings of the Fourth ACM Web Science Conference (WebSci12)*, Evanston, IL, USA, 2012, pp. 55–58.

[19] A. R. Aronson, "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program," *Proc AMIA Symp*, pp. 17–21, 2001.

[20] R. N. Kostoff, "Validating discovery in literature-based discovery," *J Biomed Inform*, vol. 40, no. 4, pp. 448–450, Aug 2007.

[21] R. N. Kostoff, M. Briggs, J. Solka, and R. Rushenberg, "Literature-related discovery (LRD): Methodology," *Technological Forecasting and Social Change*, vol. 75, no. 2, pp. 186–202, Feb. 2008.

[22] R. N. Kostoff, "Where is the discovery in literature-based discovery?" in *Literature-based Discovery*, ser. Information Science and Knowledge Management, P. Bruza and M. Weeber, Eds. Springer Berlin Heidelberg, 2008, vol. 15, pp. 57–72.

[23] Q. T. Zeng, T. Tse, G. Divita, A. Keselman, J. Crowell, A. C. Browne, S. Goryachev, and L. Ngo, "Term identification methods for consumer health vocabulary development," *J. Med. Internet Res.*, vol. 9, no. 1, p. e4, 2007.