EpiphaNet: An Interactive Tool to Support Biomedical Discoveries

Journal:	*AMIA 2010 Annual Symposium*
Manuscript ID:	Draft
Manuscript Type:	Paper
Date Submitted by the Author:	
Complete List of Authors:	Cohen, Trevor; UTHealth, School of Health Information Sciences Schvaneveldt, Roger Whitfield, G Mukund, Kavitha Rindflesh, Thomas; National Library of Medicine

EpiphaNet: An Interactive Tool to Support Biomedical Discoveries

Trevor Cohen MBChB, PhD[1], Roger W. Schvaneveldt, PhD[2,] G. Kerr Whitfield, PhD[4] Kavitha Mukund MS[2], Thomas Rindflesch, PhD[3],
[1]University of Texas, Houston, TX; [2]Arizona State University, Phoenix, AZ, [3]National Library of Medicine, Bethesda, MD. [4]University of Arizona College of Medicine, Phoenix, AZ

Abstract

EpiphaNet is an interactive knowledge discovery system which enables researchers to explore visually sets of relations extracted from MEDLINE using a combination of language processing techniques. In this paper, we discuss the theoretical and methodological foundations of the system, and present a summary of results drawn from a qualitative analysis of over six hours of interaction with the system by basic medical scientists. The system is shown to generate associations that are both surprising and interesting within the area of domain expertise of the researchers concerned.

Introduction

The field of literature-based discovery (LBD) emerged from a therapeutically useful connection between Raynaud's phenomenon and dietary fish oil discovered by Don Swanson [1]. Since then, a number of systems that aim to support the process of knowledge discovery from the literature have been developed and evaluated [2]. Several of these systems are accessible through web-based interfaces to provide scientists with access to methodological advances in the field [3]. However, with the occasional exception [4], very few studies of the use of such systems by scientists exist in the literature. Furthermore, there is evidence that Swanson's ideas, while well received by the information and library science community, have not penetrated significantly into the biomedical research community [5].

This paper presents the theoretical and methodological underpinnings of EpiphaNet (http://epiphanet.uth.tmc.edu), a tool that harnesses the computational power of recent advances in distributional semantics to encourage innovation by allowing scientists to explore associations they would not otherwise encounter. In addition, we include as proof-of-concept summary statistics and illustrative excerpts from a detailed qualitative analysis of over six hours of interaction between basic medical scientists and the EpiphaNet system.

EpiphaNet

EpiphaNet is motivated by abductive reasoning as proposed by C. S. Peirce, who maintained that an understanding of scientific reasoning requires an analysis of the process through which new hypotheses are discovered [6]. Drawing inspiration from Peirce's work is not without precedent in LBD: Bruza and his colleagues make a persuasive argument for the need for a cognitively and computationally economical method through which new hypotheses might be identified in order to model abduction [7]. Following this argument, EpiphaNet aims to provide an extension of the associative process through which plausible hypotheses are generated, by providing scientists with access to associations derived from the breadth of MEDLINE using contemporary computational and language-processing techniques.

EpiphaNet learns general associations between concepts using Reflective Random Indexing (RRI) [8], a recently emerged variant of the Random Indexing (RI) method [9] that improves on its ability to derive associations between terms that do not co-occur directly [9]. RRI is used to identify general associations between concepts extracted from MEDLINE by the MetaMap system [10]. In addition, RRI has the desirable property of being able to map statistically between UMLS concepts and terms based upon their distribution across a common set of documents. Consequently, in addition to providing general associations between UMLS concepts, RRI is used to map statistically between terms and UMLS concepts allowing searchers to specify a concept with terms of their choice, and select the UMLS concept that best fits the idea in mind. Figure I illustrates the five nearest neighbors of a search on the term “merlot”, using EpiphaNet's RRI-based “Translate” feature which maps between terms and UMLS concepts..

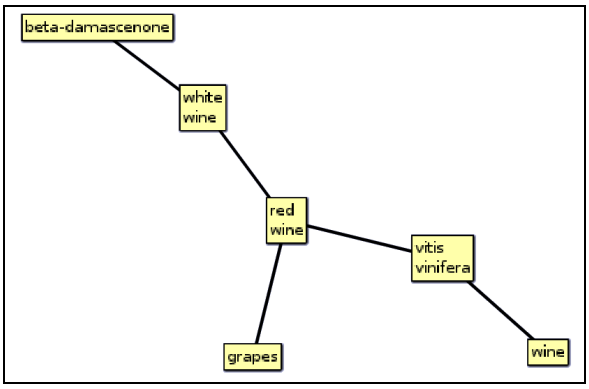

Figure 1: General associations of the term "merlot" (from the current web-based edition of EpiphaNet).

In addition to general associations, EpiphaNet provides access to more than twenty million predications (also known as object relation object triplets, for example "merlot IS-A wine") extracted by the SemRep system [11] from MEDLINE abstracts and titles added to MEDLINE over the past decade. These predications are also mapped to a compact vector space representation, in this case using Predication-based Semantic Indexing [12].

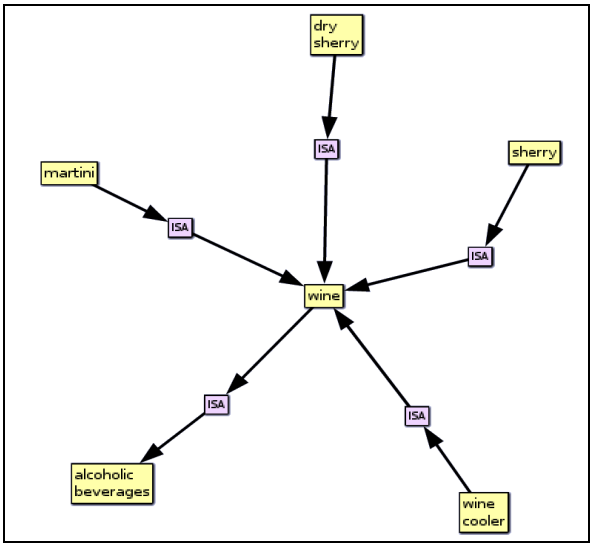

Figure 2: Specific associations of the UMLS concept "wine".

Consequently, both general co-occurrence based associations and specific predication-based associations extracted from MEDLINE over the past decade are encoded in compact in-RAM vector spaces. In both cases, these vector spaces encode distributional information from across this corpus, and consequently represent knowledge beyond the scope of what is likely to be absorbed any human reader. Figure II illustrates the five nearest predication-based associations of the UMLS concept wine (MEDLINE is surprisingly well-informed on the subject).

EpiphaNet also allows users to specify what the sort of predication relations that are of interest, by limiting these to a template of biological (for example "associated_with") or clinical (for example "treats") relations based on templates used in previous research. This can lead to quite different results when searching on the same concept. For example, the strongest predication relationships for the UMLS concept "asthma" are "positive skin test ASSOCIATED_WITH" and "alvesco TREATS" when restricting to biological and clinical predications respectively. As is evident in Figures I and II, the results of an EpiphaNet search consist of a network of associated UMLS concepts. As all concept vectors have a measurable association to one another, Pathfinder network scaling [13] is used to identify the most significant associations between concepts within this small network of neighbors. Pathfinder has been shown to generate intuitively interpretable networks of association based on human judgments of conceptual association [13], and as such is a natural choice for the display of associations derived from across the MEDLINE database. Once a network is generated, on screen layout is achieved using a force-directed algorithm provided by the Prefuse visualization library [14], written in Java. However, the current version of EpiphaNet is browser based, and consequently uses the Actionscript-based Flare visualization library (http://flare.prefuse.org/) instead.

EpiphaNet provides users with the option to investigate associations of interest by searching the Pubmed and OMIM databases after clicking on concepts within the network. In addition, when a search includes more than once concept, the system will attempt to link these concepts by finding near neighbors to a combined vector representation that includes both of them. This sort of search is similar in nature to "closed" LBD, in which a linking term between two terms that do not co-occur in the literature is sought, and can be performed with or without predications. A predication-based search linking two concepts is illustrated in Figure III. The path involves more than one intermediate concept: "prozac" is linked via an "isa" link to its generic name, "fluoxetine", which is then joined to "amitryptiline" by their shared relation "isa" antidepressant. This ability to join concepts with pathways spreading across multiple links is afforded

by the ability to combine the vector representations for several concepts into a composite vector, which can then be used to search through the remaining concepts for strongly associated links.

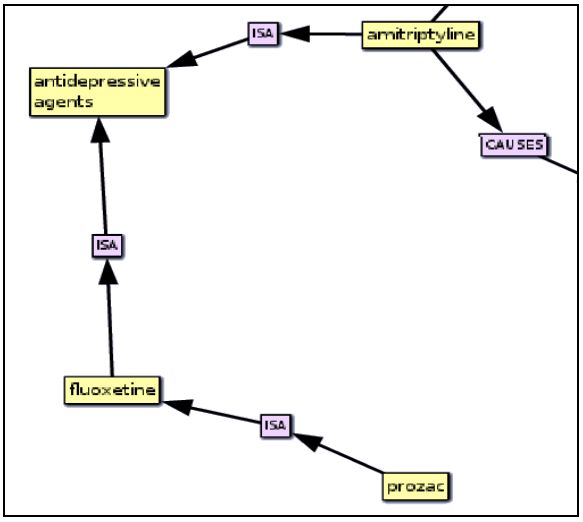

Figure 3: Excerpt of network linking search concepts "amitryptiline" and "prozac".

Evaluation

In the following section we present a summary of, and excerpts from a total of 10 sessions of between 20 and 51 minutes in duration, in which two subjects interacted with the EpiphaNet system while pursuing hypotheses of interest. In nine of these sessions, the subject was author G. K.W., a domain expert in the molecular biology of Vitamin D. In one session, numbered session eight the subject was an advanced undergraduate student with similar interests. All sessions were captured using audio- and screen-recording with the open source tool, RecordMyDesktop, transcribed to text, and then analyzed using the N-Vivo suite of software for qualitative data analysis. For the purpose of this paper, we will focus our discussion on observable evidence that EpiphaNet facilitates the discovery of associations that are new to the subject, as well as presenting an analysis of excerpts that illustrate this process in the context of an EpiphaNet session.

Results and Discussion

As illustrated in Table I, the record sessions present some evidence that EpiphaNet does support the recognition of associations that are new to the subject, and on average around nine such associations were identified in each session. This is remarkable given that most of these sessions focused on topics within the domain of expertise of the researcher concerned. Frequently (around six times per session on average), the subject would choose to investigate an association of interest in further detail using either PubMed search, or OMIM search, which was added as a feature once it was observed that this was a frequently used resource.

Session Number	Duration	Surprising Associations	Pubmed or OMIM	Printed Abstracts	Note Taking
1	20 min	7	3	0	0
2	35 min	12	6	0	3
3	36 min	6	4	2	0
4	35 min	2	1	1	0
5	45 min	15	9	4	1
6	42 min	7	5	1	0
7	31 min	13	6	2	1
8	43 min	2	2	1	0
9	51 min	16	13	2	0
10	47 min	11	9	2	3
TOTAL	~6.5 hrs	91	58	15	8
MEAN	39 min	9.1	5.8	1.5	0.8

Table I: Summary of Recorded Sessions

The series of excerpts that follow are drawn from the 10 recorded sessions, illustrate the ways in which EpiphaNet has been shown to mediate the discovery of surprising associations within the user's domain of expertise. Each excerpt is followed by either a comment by the subject of the experiments (*italics*) or a description of the significance of the excerpt.

Excerpt 1: "Oh! Look at this... prostate specific membrane antigen...I didn't know that...2008.. I didn't know that. There are lot of papers but this one they have actually found that it regulates one of the genes, wow! Ok ok this is the pubmed search. Wow! I wonder if they have a...a... response element "

Negative regulation of gene expression by VDR is generally more complicated than positive regulation and this gene is no exception - this report (Serda et al 2008) will become part of our database for negative regulation - we are currently pursuing negative regulation of several genes by VDR, one of which is S100A8 = MRP8.

Excerpt 2: "this is the... receptor for advanced glycosylation end products and s100a8 and s100a9 play an essential role in colitis carcinogenesis..Wow! This is really new... these two proteins somehow interact...this is a great set of articles. "
This novel association (that s100a8 and s100a9 are involved in colon carcinogenesis) could become a focus of future research by one of our collaborators who is interested in the connection between vitamin D and colon cancer.

Excerpt 3: "FGF5...and this is a prime example of fibroblast growth factor...as they stimulate...4 independent mutations. We are very interested in genes that stimulate hair cycle... the vdr in hte hair less gene we are studying has tremendous effect but don't know what the mechanism. FGF18 is capable of inducing"
The possible relation between FGF18 and the hair cycle is very interesting to us since we have already shown that FGF23, a gene related to FGF18, is the target of regulation by VDR. We also know that VDR controls the hair cycle but target genes are not known that might mediate this FGF18 might be a candidate for one of those target genes for VDR since we already know that a related gene, FGF23, is regulated by VDR.

Excerpt 4: a synonym for S100A8 or calgranulin-MRP8. Oh!I can do SIN 1A- MRP8. Oh there is another synonym Oh yeah sorry..I am trying to make sense out of this..it doesn't seem to be the same synonym MRP8.I swear..that this is a..I swear that MRP8 is a synonym. Or is it MPR8?..MRP! "

I believe that the fact that S100A8 and MRP8 are synonyms was revealed to me by an OMIM search prompted by Epiphanet (perhaps an unrecorded session) This knowledge allowed me to find a key paper in PubMed showing that MRP8/s100A8 is regulated by 1,25D This is now the subject of some of our experiments in the lab (confirming this regulation and searching for vitamin D response elements that mediate this regulation.

Excerpt four illustrates the nomenclature issues that at times presented problems both for the specification and the interpretation of EpiphaNet searches. In this case, the issue had a positive effect as the subject was able to identify an unexplored body of literature after identifying a previously unrecognized synonym for a gene of interest.

Excerpt 5: "Maybe I should try VDR...oh! I remember from last time...Vitamin D receptor is what it wants to do These are two very related concepts...That's an analog...NCOR1 gene...it's a nuclear co activator and it works with several steroid receptors...nuclear co activator...nuclear co repressor...I better look it up..."

Our lab had concluded based on a preliminary experiment that NCOR1 did NOT interact with VDR. However, Epiphanet (and associated PUBMED searches) identified several recent papers that contradict this conclusion. I was not aware of these more recent reports and would not have searched for them had I not been alerted by Epiphanet that our conclusion was incorrect and had been superceded by newer data.

Excerpt 6: "TOB interacts with Vit D receptor..Wow! This is the 2008... TOB2...I am learning a lot here...TOB2 and this is in stromal cells...FEBS letters. That's a journal that I don't always look into...it's a very general experimental biology journal...and doesn't specialize in my area."

This excerpt illustrates the way in which EpiphaNet encourages researchers to evaluate associations drawn from outside the sphere of their usual literature review. In another instance, the subject traced a novel association to the dental literature.

Excerpt 7: "Prolactin aspartate 179... I didn't know that... 179 diminishes the effect of UV light.. now UV light is what forms vitamin D in the skin. …now what does that *[a Prolactin mutant]* have to do with the Vitamin D receptor? I want to see what it has to do with both of those. Now I am doing a PubMED search with both of those... Vitamin D and this mutant form of prolactin. Now this is completely unexpected... it's completely new to me. Although the papers are '05 and '07. Wow! That's weird! This is in human prostate cancer cells... *[and I recall that]* 1,25 dihydroxy*[vitamin D3]* is known to have anticancer effects."

This relationship was new to the subject, who proceeded to follow up with a PubMED search, which can be accomplished from within the EpiphaNet interface. The new information obtained from the abstracts resulting from the search is then integrated with the subject's own knowledge related to the anti-cancer effects of Vitamin D3.

Some limitations of EpiphaNet were also apparent in the recorded sessions. In particular, on many occasions subjects' encountered issues with nomenclature in which a term used as a search cue had other unintended meanings. This problem also manifested upon examination of search results, as on several occasions subjects were not able to interpret the UMLS concept presented. This problem has been alleviated to some extent by the addition of OMIM search, given that OMIM entries usually contain a list of synonyms. As it has evolved, EpiphaNet has

departed somewhat from the traditional model of LBD: in early iterations the system offered users the ability to perform searches for related concepts that do not appear together in the literature, the primary concern of most LBD systems. However, based on testing with a research scientist, this feature was removed, as it became quickly apparent that our subjects were more motivated by discovering connections that were "new to them" than "new to science". This raises an interesting issue for the developers of LBD systems. The theoretical and methodological issues associated with identifying previously undiscovered and meaningful associations present an interesting and appealing challenge for information scientist. However, as observed by Smalheiser and his colleagues [15], the intended users LBD systems do not draw clear distinctions between the tools of information retrieval and the formal methodologies of LBD. Consequently we have shifted the focus of the system toward providing a dynamic and interactive visual presentation of associations extracted from the literature. This represents a deliberate move on our part away from the goal of fully automated knowledge discovery, and toward the goal of a system that supports the distributed cognitive process of knowledge discovery in a dynamic and responsive manner. As illustrated by the excerpts presented in this paper, EpiphaNet is able to achieve this goal to some extent – subjects' exploration and discovery of new relations is tightly integrated with their exploration of the literature and integration of their own expert knowledge.

Conclusion

EpiphaNet is a novel tool that uses emerging methods of distributional semantics to support an interactive visual display of relations between UMLS concepts. Qualitative studies of users suggest the tool is able to support the discovery of associations that are novel and surprising to scientists within their domain of expertise. The tool is currently available online, and future evaluation efforts will focus on the analysis of the activities of a wider range of potential users.

References

[1] Swanson DR. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspect Biol Med. 1986;30(1):7-18.

[2] Sehgal AK, Qiu XY, Srinivasan P. Analyzing LBD Methods using a General Framework. Literature-based Discovery. 2008;:75.

[3] Weeber M, Kors JA, Mons B. Online tools to support literature-based discovery in the life sciences. Briefings in bioinformatics. 2005;6(3):277-286.

[4] Skeels MM, Henning K, Yildiz MY, Pratt W. Interaction design for literature-based discovery. ACM New York, NY, USA; 2005. p. 1785-1788.

[5] Spasser MA. The enacted fate of undiscovered public knowledge. JASIS. 1997;48(8):707-717.

[6] Peirce CS. Abduction and Induction. In: J. Buchler (Ed.) Philosophical writings of Peirce. New York: Routledge; 1940.

[7]Bruza P, Cole R, Song D, Bari Z. Towards Operational Abduction from a Cognitive Perspective. Logic Jnl IGPL. 2006 Mar;14:161-177.

[8] Cohen T, Schvaneveldt R, Widdows D. Reflective Random Indexing and Indirect Inference: A Scalable Method for Discovery of Implicit Connections. JBI. 2009 Sep 14 [epub ahead of print].

[9] Kanerva P, Kristofersson J, Holst A. Random indexing of text samples for latent semantic analysis. Proceedings of the 22nd Annual Conference of the Cognitive Science Society. 2000;1036.

[10] Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. 2001;17:17-21.

[11] Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. Journal of Biomedical Informatics. 2003;36:462-477.

[12] Cohen T, Schvaneveldt R, Rindflesch T. Predication-based Semantic Indexing: Permutations as a Means to Encode Predications in Semantic Space. Proc AMIA symp, 2009.

[13] Schvaneveldt RW. Pathfinder associative networks: studies in knowledge organization. Ablex Publishing Corp. Norwood, NJ, USA; 1990.

[14] Heer J, Card SK, Landay JA. prefuse: a toolkit for interactive information visualization. Conference on Human Factors in Computing Systems. 2005;:421-430.

[15] Smalheiser NR, Torvik VI, Bischoff-Grethe A, Burhans LB, Gabriel M, Homayouni R, Kashef A, Martone ME, Perkins GA, Price DL, others. Collaborative development of the Arrowsmith two node search interface designed for laboratory investigators. Journal of biomedical discovery and collaboration 2006;1(1):8.