# SemanTex: Semantic Text Exploration Using Document Links Implied by Conceptual Networks Extracted from the Texts*

Suad Aldarra[1], Emir Muñoz[1], Pierre-Yves Vandenbussche[1], and Vít Nováček[2]

[1] Fujitsu (Ireland) Limited
Airside Business Park, Swords, Co. Dublin, Ireland
E-mail: Firstname.Lastname@ie.fujitsu.com
[2] Insight @ NUI Galway (formerly known as DERI)
IDA Business Park, Lower Dangan, Galway, Ireland
E-mail: vit.novacek@deri.org

## 1 Introduction

Despite of advances in digital document processing, exploration of implicit relationships within large amounts of textual resources can still be daunting. This is partly due to the 'black-box' nature of most current methods for computing links (*i.e.*, similarities) between documents (*c.f.*, [1] and [2]). The methods are mostly based on numeric computational models like vector spaces or probabilistic classifiers. Such models may perform well according to standard IR evaluation methodologies, but can be sub-optimal in applications aimed at end users due to the difficulties in interpreting the results and their provenance [3, 1].

Our Semantic Text Exploration prototype (abbreviated as SemanTex) aims at finding implicit links within a corpus of textual resources (such as articles or web pages) and exposing them to users in an intuitive front-end. We discover the links by: (1) finding concepts that are important in the corpus; (2) computing relationships between the concepts; (3) using the relationships for finding links between the texts. The links are annotated with the concepts from which the particular connection was computed. Apart of being presented to human users for manual exploration in the SemanTex interfaces, we are working on representing the semantically annotated links between textual documents in RDF and exposing the resulting datasets for particular domains (such as PubMed or New York Times articles) as a part of the Linked Open Data cloud.

In the following we provide more details on the method and give an example of its practical application to browsing of biomedical articles. A video example of a specific SemanTex prototype to be demonstrated at the conference can be looked up at http://goo.gl/zL8lJ2.

* This work has been supported by the 'KI2NA' project funded by Fujitsu Laboratories Limited in collaboration with Insight @ NUI Galway.

## 2 Method

**Extracting Conceptual Networks.** For extracting links between concepts in the texts we use methods we introduced in [4]. The essentials of the method are as follows: (1) Extracting noun phrases that may refer to domain-specific concepts (using either a shallow parser for general texts or biomedical named-entity recognition tool for life sciences). (2) Computing co-occurrence relationships between the extracted noun phrases by means of point-wise mutual information (PMI). (3) Filtering out the relationships with the PMI scores below a threshold. (4) Computing (*cosine*) similarity relationships based on the co-occurrence ones.

**Computing Paths between Documents.** From a conceptual network, one can generate sets of paths leading out from every concept. To prevent a combinatorial explosion, we limit the paths by two factors: (1) the maximum path length; (2) the minimum product of the edge weights of the path. From the set of such paths associated with particular nodes, paths between the original documents (*i.e.*, text-to-text links semantically annotated by the concepts appearing on them) can be generated using inverted indices of concept-text provenance. For instance, imagine a text $A$ contains concept $x$. Now assume that $x$ is related to a concept $y$ in text $B$ via a path $(x, u, v, y)$. Then we can say the texts $A$ and $B$ are related by a $(x, u, v, y)$ path.

**Selecting the Most Relevant Paths.** The critical part of the method is finding out which paths are most promising out of potentially huge numbers of them. For that we use multi-objective optimisation of several specific complexity, coherence and entropy measures introduced in [4]. We follow certain intuitive assumptions when selecting the path measures to optimise: (1) Paths leading through more complex environs are more informative for a user. (2) Paths surrounded by many highly balanced (*i.e.*, entropic) topics are more informative. (3) Coherent paths with gradual topical changes on the way are better (less chaotic, more focused progression from one topic to another en route to the linked text). (4) It is more interesting, considering an Information Retrieval point of view, when one ends up in a topically distant (incoherent) area (once the progress through the topics is gradual, *i.e.*, less random). The result of this step is a set of optimal (non-dominated) text-to-text paths that can be further ranked according to their combined score.

## 3 Usage

To demonstrate the SemanTex technology, we have applied it to the corpus of Parkinson's Disease article abstracts from PubMed which we experimented with in [4]. As can be seen in Figure 1, the front-end has been incorporated into a PubMed look-and-feel. Domain-specific concepts have been highlighted in the abstract display (the darker the shade, the more the concept is important for the given abstract). After clicking on any of the highlights, a separate 'Path Diagram' window is displayed where one can navigate paths leading from the selected concept. The nodes on the paths can be expanded with further connections

while the corresponding related articles are always displayed in the bottom of the window. Clicking on a related article leads to the article view. One can also explore articles related by a path to the currently browsed one. A diagram of paths that connect the articles via the concepts in them can be displayed as well.

Figure 1 illustrates SemanTex on an example of article about the correlation of caffeine consumption, risk of Parkinson's disease and the related differences between men and women. When exploring the concept *'high caffeine consumption'*, one can continue to the *'hormones'* and *'women'* nodes. Expanding the *'women'* link shows many concepts related to women's health, such as *'oophorectomy'* (removal of ovaries). There is a single article related to that concept, dealing with increased risk of parkinsonism in women who underwent oophorectomy before menopause. This shows how one can quickly explore a problem from many different viewpoints with SemanTex, linking an article dealing with the influence of particular hormonal levels on the development of Parkinson's Disease in women with another article looking into higher risk of parkinsonism due to lower levels of estrogen caused by the pre-menopausal removal of ovaries.

When further exploring some articles related to the last one, one can see all the paths that connect them. For instance, a study about Dutch elderly people is linked to the oophorectomy article by means of four paths, all involving concepts clearly related to common geriatric ailments. This illustrates the possibility of smooth topical progression in exploring the articles.

## 4 Conclusions

In this work, we have presented SemanTex, an application that discovers implicit, semantically annotated links within a corpus of textual resources. We have implemented a sample prototype of the technology deployed on PubMed articles using the standard PubMed look-and-feel. This was to show how we can easily add value to many traditional applications involving exploration of large numbers of textual resources. In a similar way, we will implement SemanTex versions for New York Times and Wikipedia articles within an evaluation trial of the technology. Last but not least, we plan to generate RDF representations of the semantically annotated links between texts computed by particular instances of SemanTex and expose them as a part of the LOD cloud.

## References


1. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press (2008)
2. Lin, J., Wilbur, W.J.: PubMed related articles: a probabilistic topic-based model for content similarity. BMC Bioinformatics **8**(1) (2007)
3. Grefenstette, E.: Analysing document similarity measures. Master's thesis, University of Oxford (2009)
4. Nováček, V., Burns, G.A.: SKIMMR: Facilitating knowledge discovery in life sciences by machine-aided skim reading. PeerJ (2014) In press, see `https://peerj.com/preprints/352/` for a preprint.

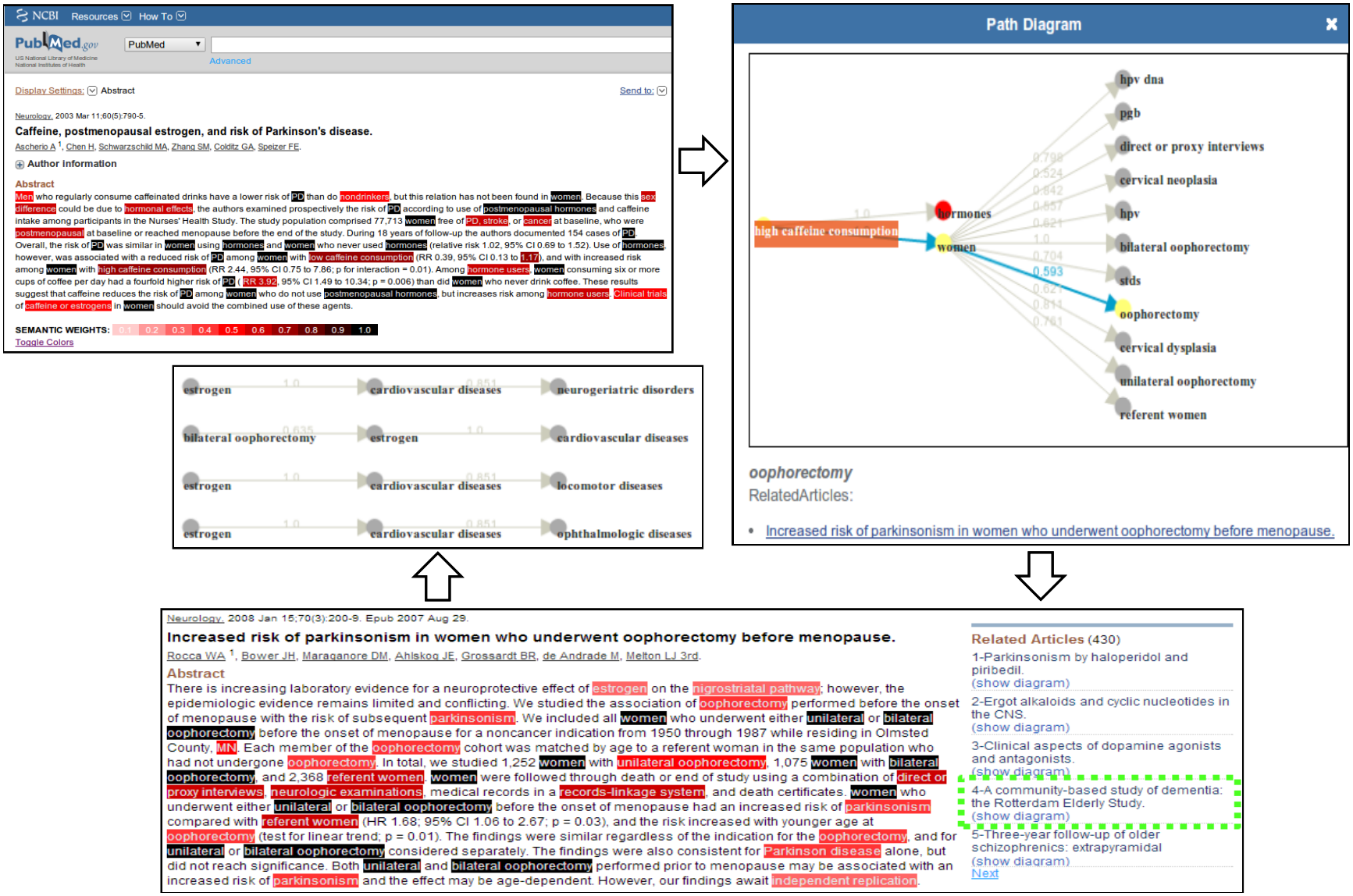


**Fig. 1.** SemanTex usage example - Parkinson's Disease