Curr Cardiovasc Risk Rep (2013) 7:1–8
DOI 10.1007/s12170-012-0288-3

Text Mining Supporting Search for Knowledge Discovery in Diabetes

Sophia Ananiadou · Tomoko Ohta ·
Martin K. Rutter

Published online: 22 December 2012

Abstract Due to increasing specialization, silo effects and literature deluge, researchers are struggling to draw out general truths and to generate testable hypotheses. This is especially true when considering the needs of biomedicine. Medicine faces many challenges, not least the fragmentation into multiple subspecialist areas, and, at the same time, the need for cutting-edge research to be interdisciplinary. There are also issues of communication and understanding between those working at different '-omics levels' and those working in a myriad of diverse areas including: basic research, translational medicine, clinical care, clinical trials, epidemiology, public health, clinical guideline development, evaluation of new drugs and treatments and personalized medicine. Most importantly, there is a lack of effective communication between these groups and members of the general public who seek to become better informed about their health. Different people have different views, perspectives and information needs relating to the same topic. Text mining methods can support information access for diverse groups such as researchers, clinicians, caregivers, patients and also members of the general public.

S. Ananiadou (✉) · T. Ohta
National Centre for Text Mining, University of Manchester,
Manchester M1 7DN, UK
e-mail: Sophia.Ananiadou@manchester.ac.uk

S. Ananiadou · T. Ohta
School of Computer Science, University of Manchester,
Manchester, UK

M. K. Rutter
Endocrinology and Diabetes Group, Developmental Biomedicine
Research Group, University of Manchester, Manchester, UK

M. K. Rutter
Manchester Diabetes Centre, Manchester Royal Infirmary,
Manchester, UK

Keywords Text mining · Knowledge discovery · Diabetes · Biomedical text mining

Introduction

Due to their multidimensional nature, diseases such as diabetes present an excellent case study for using sophisticated analysis of text based on meaning (semantics) [1]. The understanding of complex diseases such as diabetes requires the discovery of entities and facts from several domains including: clinical observation, epidemiology, genomics, transcriptomics, epigenetics, proteomics and metabolomics [2]. The evidence required to generate innovative ideas and hypotheses for use in research, diagnostics and therapy is often hidden in the text of full articles. This is especially true in biomedicine where researchers and clinicians have access to several online textual resources such as abstract or full-text of scientific literature, clinical data and clinical trial reports. For example, the US National Library of Medicine's PubMed (http://www.ncbi.nlm.nih.gov/pubmed/) is a primary search facility for biomedical literature consisting of more than 22 million citations, and is growing in size by more than half a million new abstracts every year. In PubMed, a simple query such as 'diabetes' yielded, at the time of writing, 419,960 results and 'diabetes AND cardiovascular risk' yielded 36,151 results. Unfortunately, current information retrieval engines do not provide rapid and accurate answers to specific queries [3]. Given the rapid growth of the literature, there is an urgent need for text mining tools that improve the accuracy and specificity of information retrieval and at the same time avoid data overload for users. Here we present promising new ways to achieve this by searching based on semantic metadata derived from text mining technologies.

Biomedical Text Mining

Given the overwhelming amount of biomedical knowledge recorded in textual form including full papers, abstracts and grey literature, there is a need for techniques that can help users not only to identify relevant knowledge, but also to extract, manage and integrate this information [4–6]. Some text mining tools are already available that enrich biomedical papers with semantic annotations for richer querying and also to extract relations between annotated entities. For example, it is now possible to identify proteins within text, and pull out their physical interactions and associations with disease states, phenotypes and other terms [7, 8]. One can also expand on and disambiguate biomedical abbreviations, find their synonyms and map them to manually curated databases. For example, the acronym 'APP' expands to 26 different definitions, such as 'amyloid precursor protein', 'acute phase proteins', 'aminopeptidase' and many others. Text mining tools such as Acromine[1,2] help us to expand and disambiguate biomedical acronyms, a very prolific type of synonym creation [9, 10]. In addition, we can create semantically rich queries over the literature, such as 'what activates p53?' which could provide a more meaningful search than simply requesting information using the search term 'p53 activation' [10, 11]. Owing to the increasing volume and rate of scientific publication, it is clear that the automated processes offered through text mining could be a major resource facilitating the understanding of the biomedical literature.

However, the plethora of techniques currently available still focus on textual co-occurrences, using a 'bag-of-words' approach and similar shallow techniques, which cannot provide meaningful relations among entities and terms needed to identify complex relations for disease such as causality [12]. Text retrieval systems such as NLM's PubMed permit only Boolean (AND/OR) combinations of search terms.

Other biomedical text mining systems such as iHOP,[3] CiteXplore[4] and MedlineRanker[5] go a step further than conventional search systems in that they include protein/gene recognition and protein–protein interaction extraction, but they do not use relations or events for searching.

To help elucidate, for example, the roles played by biomolecules in important biological processes and disease, text mining systems have to tackle the complex problem of extracting and identifying the context and type of such relationships. The regulation of biological processes is crucial to control and maintain the life cycles of organisms [12]. A bioprocess may consist of any number of chemical reactions or other types of biological events that may result in maintenance, changes or transformations of the organism. Scientists often need to gather scientific evidence of how potential molecular targets are related, by observation or intervention, to pathophysiological processes or disease. Typically, this involves collecting evidence from cell lines, model organisms and then from clinical samples [13]. However, this process is very expensive and time consuming. To avoid inefficient and sometimes unnecessary duplication of work, scientists could first review all prior research activity in their area of interest to identify the unanswered questions and the testable hypotheses. Laboratory resources could then be directed more efficiently to explore those novel questions.

Text mining results can be either too noisy or too restricted to be useful if they don't identify the underlying mechanisms of biological processes. Therefore, more advanced analytical methods are necessary that undertake deeper semantic analysis. Semantic searches incorporate the context of the search, variation in use of words including synonyms, concept matching (e.g. mechanisms) and the ability to generate general and specific queries. To achieve these aims, text miners have developed techniques that automatically extract events of biomedical relevance pertaining to processes such as protein–protein interactions, and protein–disease and disease-disease associations from the literature as described above. However, for more advanced systems answering semantic and focused questions, it is essential to recognize mechanisms relevant to disease (e.g. activate, phosphorylate, bind, inhibit) as well as named entities (e.g. protein, disease, metabolites) [14].

In the following sections, we demonstrate how advanced text analytics that undertake deeper semantic analysis can support knowledge discovery and hypothesis generation. The text mining services described here have been developed by the UK's National Centre for Text Mining, and are freely available.[6]

Text Mining Services

Searching Using Semantic Types

KLEIO[7] is an advanced information retrieval system that offers semantic searches across MEDLINE abstracts using 'semantic metadata' derived from text mining. Semantic metadata improve knowledge capture and search by adding multiple layers of annotation to literature such as named entities, relations/events but also extralinguistic knowledge, e.g. causality. Semantic searches are carried out using

[1] http://www.nactem.ac.uk/software/acromine/

[2] http://www.nactem.ac.uk/software/acromine_disambiguation/

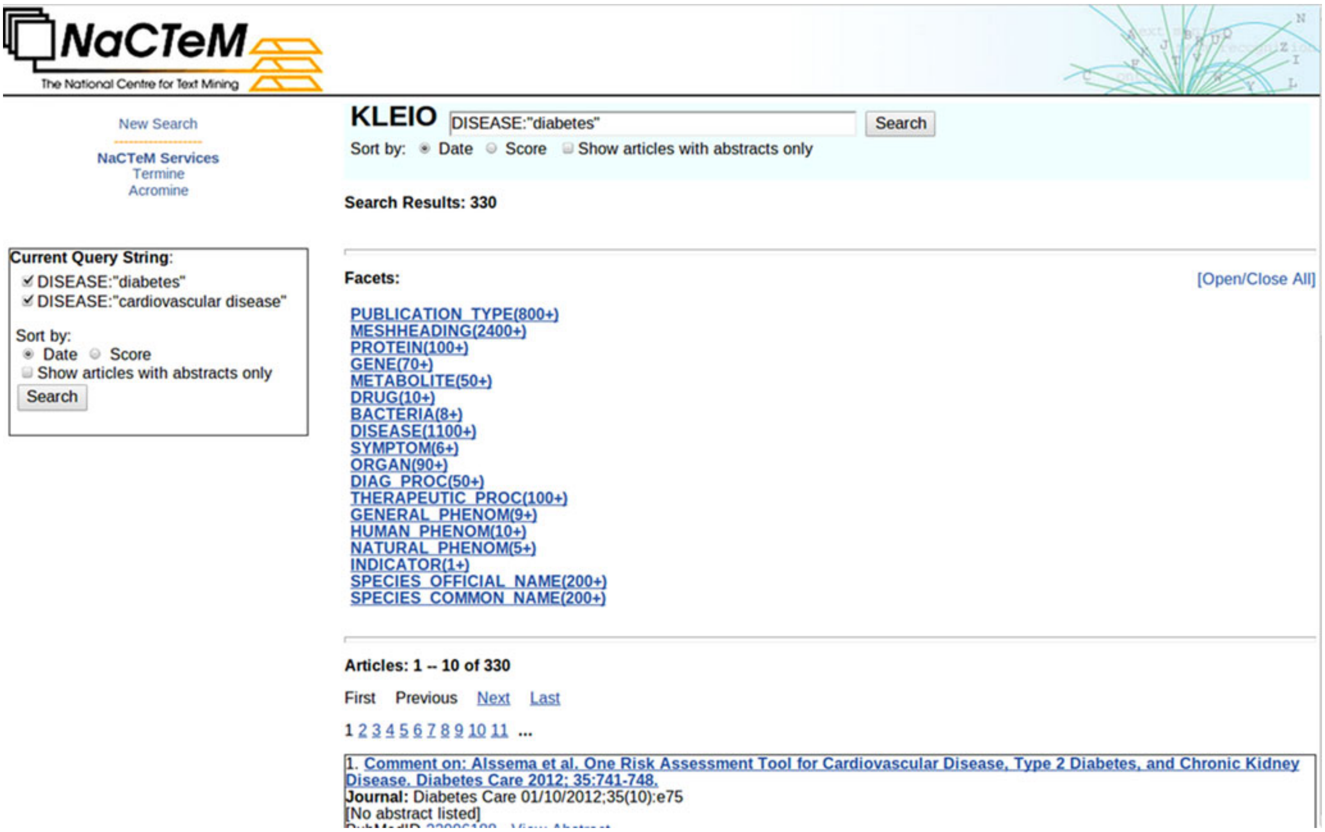
[3] http://www.ihop-net.org/UniPub/iHOP/

[4] http://www.ebi.ac.uk/citexplore/

[5] http://cbdm.mdc-berlin.de/tools/medlineranker/#

[6] http://www.nactem.ac.uk/services.php

[7] http://www.nactem.ac.uk/Kleio



New Search

NaCTeM Services
Termine
Acromine

KLEIO DISEASE:"diabetes" Search
Sort by: ● Date ○ Score □ Show articles with abstracts only

Search Results: 330

Current Query String:
☑ DISEASE:"diabetes"
☑ DISEASE:"cardiovascular disease"

Sort by:
● Date ○ Score
□ Show articles with abstracts only
Search

Facts: [Open/Close All]

PUBLICATION_TYPE(800+)
MESHHEADING(2400+)
PROTEIN(100+)
GENE(70+)
METABOLITE(50+)
DRUG(10+)
BACTERIA(8+)
DISEASE(1100+)
SYMPTOM(6+)
ORGAN(90+)
DIAG_PROC(50+)
THERAPEUTIC_PROC(100+)
GENERAL_PHENOM(9+)
HUMAN_PHENOM(10+)
NATURAL_PHENOM(5+)
INDICATOR(1+)
SPECIES_OFFICIAL_NAME(200+)
SPECIES_COMMON_NAME(200+)

Articles: 1 -- 10 of 330

First Previous Next Last

1 2 3 4 5 6 7 8 9 10 11 ...

1. Comment on: Alssema et al. One Risk Assessment Tool for Cardiovascular Disease, Type 2 Diabetes, and Chronic Kidney Disease. Diabetes Care 2012; 35:741-748.
Journal: Diabetes Care 01/10/2012;35(10):e75
[No abstract listed]

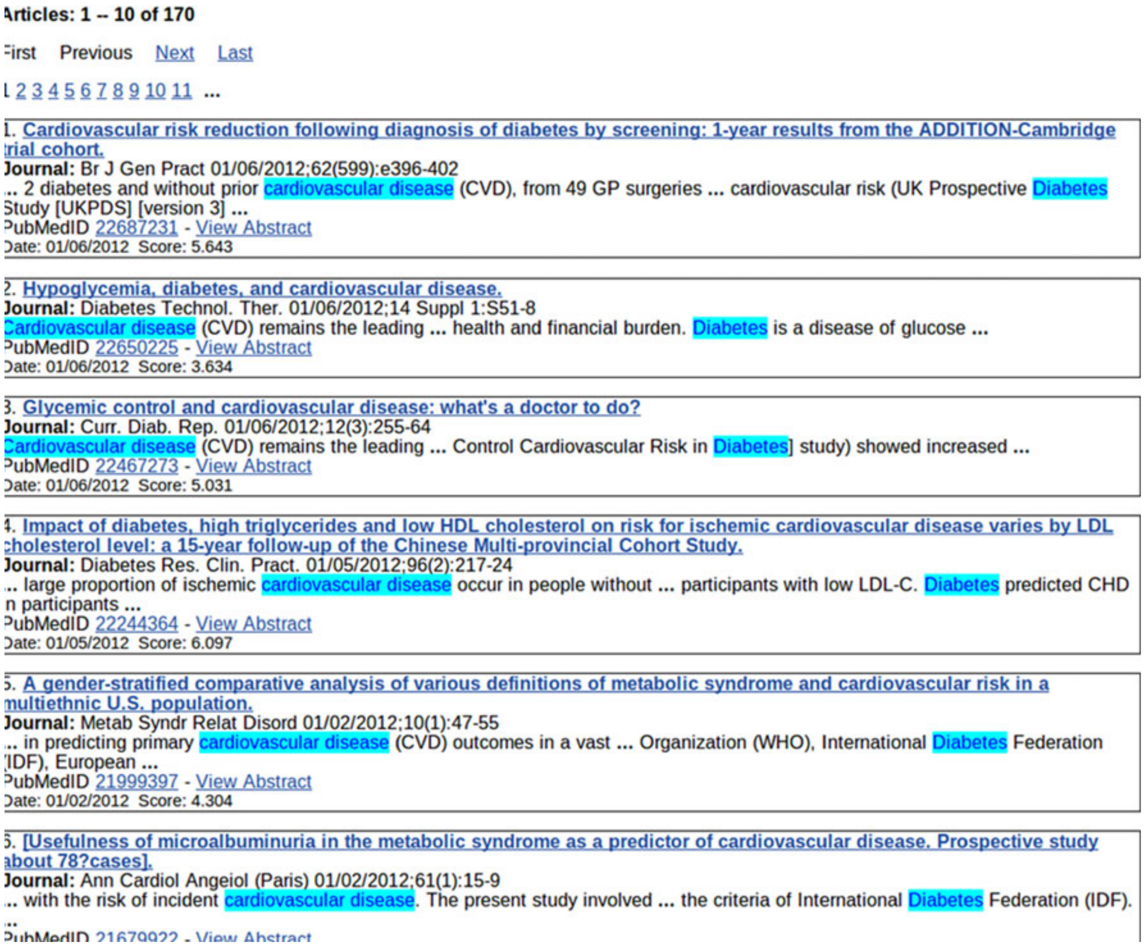
Fig. 1 Faceted search using KLEIO

techniques such as 'named entity recognition' which automatically detects and marks-up biologically important terms appearing in text, such as 'gene', 'protein', 'disease x', 'drug name' and 'metabolite name'. To improve the user experience, KLEIO provides an interactive faceted search using MEDLINE data based on semantic types. Faceted navigation allows flexible searching of metadata, and thus has been adopted by several websites. For example, in KLEIO, the user can select among different types of semantic queries suggested by the system, using a query builder. KLEIO delivers rapid responses, based on preindexed semantic types linked to synonyms, highlighting the retrieved documents along with their semantic types. Synonymy detection is handled using techniques such as 'term variability' [15] and 'normalization' [16] including 'acronym detection' and 'disambiguation'.

For example, KLEIO will link a generic query such as 'diabetes' to different forms of diabetes including 'insulin-dependent diabetes mellitus', 'type 1 diabetes mellitus', 'juvenile-onset diabetes', 'type 2 diabetes mellitus' and 'non-insulin dependent diabetes mellitus'. KLEIO will accept a query such as DISEASE:"diabetes" or a word or its acronym (e.g. 'cat', 'IL-6') or a combination of these (e.g. PROTEIN: IL-6 AND SPECIES: mus musculus). The system then retrieves all the abstracts from MEDLINE that match the query and it will show the set of 'facet labels' (semantic characteristics) that were used to construct the query (Fig. 1). The interactive navigation of search results is supported by the original criteria used by KLEIO to generate the faceted search. Clicking on a facet label causes a list of 'semantic types' or associated Medical Subject Headings (MeSH) terms to be displayed, in descending order of frequency of occurrence among the retrieved documents. By clicking on a semantic type or a MeSH term in a facet, the user can append the type or MeSH term to the semantic query, and thus narrow down the search over the current set of retrieved documents. A manageable set of useful documents can thus be retrieved in a few clicks as the query is progressively narrowed down.

Whereas PubMed keyword search gives access to 115445 abstracts for the query 'diabetes AND cardiovascular disease', KLEIO provides more focused results for a query 'DISEASE:"diabetes" AND DISEASE:"cardiovascular disease" AND MESH HEADING:"risk factors"' (Fig. 2). In the search results, each disease name 'diabetes' and 'cardiovascular disease' is recognized as a specific semantic class, and includes synonymous expressions such as 'type 2 diabetes' (Fig. 3).

Each retrieved document has a link to more detailed information. When users click the title of the document, an abstract is shown highlighting the extracted entities and also providing the metadata of the abstract with links to the original resources.



Articles: 1 -- 10 of 170

First Previous Next Last

1 2 3 4 5 6 7 8 9 10 11 ...

1. Cardiovascular risk reduction following diagnosis of diabetes by screening: 1-year results from the ADDITION-Cambridge trial cohort.
Journal: Br J Gen Pract 01/06/2012;62(599):e396-402
... 2 diabetes and without prior cardiovascular disease (CVD), from 49 GP surgeries ... cardiovascular risk (UK Prospective Diabetes Study [UKPDS] [version 3] ...
PubMedID 22687231 - View Abstract
Date: 01/06/2012 Score: 5.643

2. Hypoglycemia, diabetes, and cardiovascular disease.
Journal: Diabetes Technol. Ther. 01/06/2012;14 Suppl 1:S51-8
Cardiovascular disease (CVD) remains the leading ... health and financial burden. Diabetes is a disease of glucose ...
PubMedID 22650225 - View Abstract
Date: 01/06/2012 Score: 3.634

3. Glycemic control and cardiovascular disease: what's a doctor to do?
Journal: Curr. Diab. Rep. 01/06/2012;12(3):255-64
Cardiovascular disease (CVD) remains the leading ... Control Cardiovascular Risk in Diabetes] study) showed increased ...
PubMedID 22467273 - View Abstract
Date: 01/06/2012 Score: 5.031

4. Impact of diabetes, high triglycerides and low HDL cholesterol on risk for ischemic cardiovascular disease varies by LDL cholesterol level: a 15-year follow-up of the Chinese Multi-provincial Cohort Study.
Journal: Diabetes Res. Clin. Pract. 01/05/2012;96(2):217-24
... large proportion of ischemic cardiovascular disease occur in people without ... participants with low LDL-C. Diabetes predicted CHD in participants ...
PubMedID 22244364 - View Abstract
Date: 01/05/2012 Score: 6.097

5. A gender-stratified comparative analysis of various definitions of metabolic syndrome and cardiovascular risk in a multiethnic U.S. population.
Journal: Metab Syndr Relat Disord 01/02/2012;10(1):47-55
... in predicting primary cardiovascular disease (CVD) outcomes in a vast ... Organization (WHO), International Diabetes Federation (IDF), European ...
PubMedID 21999397 - View Abstract
Date: 01/02/2012 Score: 4.304

6. [Usefulness of microalbuminuria in the metabolic syndrome as a predictor of cardiovascular disease. Prospective study about 78?cases].
Journal: Ann Cardiol Angeiol (Paris) 01/02/2012;61(1):15-9
... with the risk of incident cardiovascular disease. The present study involved ... the criteria of International Diabetes Federation (IDF).
...
PubMedID 21679922 - View Abstract

Fig. 2 KLEIO search results

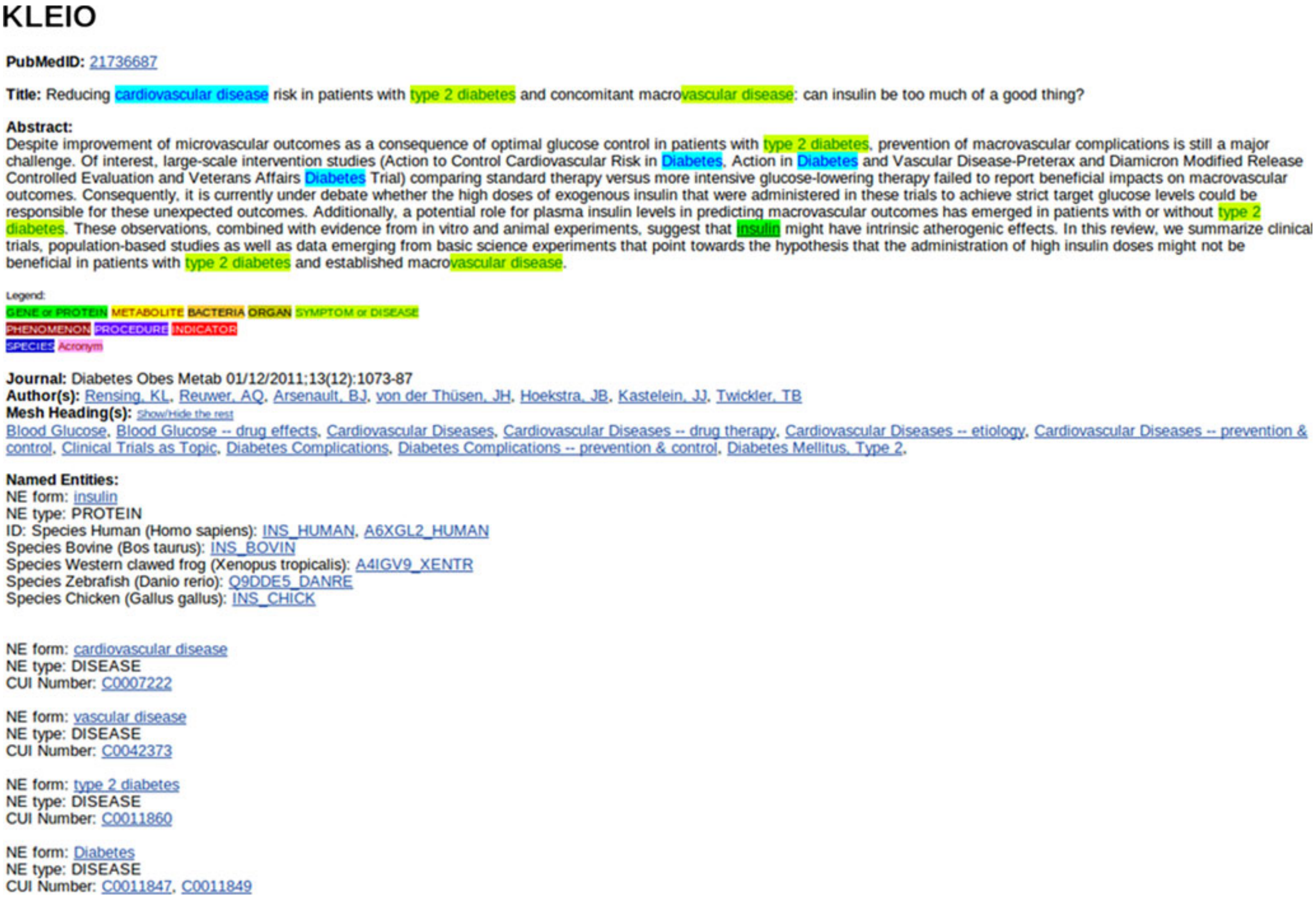
Mining Direct and Indirect Associations

FACTA+[8] is a real-time (interactive and online) text-mining system for finding and visualizing direct and indirect associations between biomedical concepts from MEDLINE abstracts. The system can be used as a text search engine like PubMed with additional features to help users discover and visualize associations between important biomedical concepts in MEDLINE abstracts retrieved by a query. Information about pair-wise associations between biomedical concepts, such as genes, proteins, diseases and chemical compounds constitutes an important part of biomedical knowledge. It is common for a researcher to ask questions such as ‘What genes are relevant to diabetes?’ or ‘What chemical compounds are relevant to diabetes?’ This novel text mining facility complements biomedical databases by providing researchers with a convenient way to find the answers to such questions from the literature.

An important feature of FACTA+ is its ability to discover indirectly associated concepts. Discovering hidden, previously unknown and potentially useful associations between biomedical concepts such as genes and diseases from the literature is a longstanding goal in biomedical text mining [17]. For example, in pioneering work, Swanson [18] and Swanson and Smalheiser [19] hypothesized the role of fish oil in the clinical treatment of Raynaud’s disease, combining different pieces of information from the literature, and the hypothesis was later confirmed by experimental evidence.

More specifically, text mining can generate new hypotheses by discovering indirect associations by combining two known associations, which are obtained from direct co-occurrence statistics. Text mining can give a probabilistic interpretation to the strengths of all novel indirect associations, which are ranked in the order of expected information quantity. For example, a common approach to automatic discovery of novel hypotheses is to combine two (or more) known associations, i.e. if concept X is associated with concept Y, and concept Y is associated with concept Z, then

[8] http://refine1-nactem.mc.man.ac.uk/facta-visualizer/



KLEIO

PubMedID: 21736687

Title: Reducing cardiovascular disease risk in patients with type 2 diabetes and concomitant macrovascular disease: can insulin be too much of a good thing?

Abstract:
Despite improvement of microvascular outcomes as a consequence of optimal glucose control in patients with type 2 diabetes, prevention of macrovascular complications is still a major challenge. Of interest, large-scale intervention studies (Action to Control Cardiovascular Risk in Diabetes, Action in Diabetes and Vascular Disease-Preterax and Diamicron Modified Release Controlled Evaluation and Veterans Affairs Diabetes Trial) comparing standard therapy versus more intensive glucose-lowering therapy failed to report beneficial impacts on macrovascular outcomes. Consequently, it is currently under debate whether the high doses of exogenous insulin that were administered in these trials to achieve strict target glucose levels could be responsible for these unexpected outcomes. Additionally, a potential role for plasma insulin levels in predicting macrovascular outcomes has emerged in patients with or without type 2 diabetes. These observations, combined with evidence from in vitro and animal experiments, suggest that insulin might have intrinsic atherogenic effects. In this review, we summarize clinical trials, population-based studies as well as data emerging from basic science experiments that point towards the hypothesis that the administration of high insulin doses might not be beneficial in patients with type 2 diabetes and established macrovascular disease.

Legend:
GENE or PROTEIN METABOLITE BACTERIA ORGAN SYMPTOM or DISEASE
PHENOMENON PROCEDURE INDICATOR
SPECIES Acronym

Journal: Diabetes Obes Metab 01/12/2011;13(12):1073-87
Author(s): Rensing, KL, Reuwer, AQ, Arsenault, BJ, von der Thüsen, JH, Hoekstra, JB, Kastelein, JJ, Twickler, TB
Mesh Heading(s): Show/Hide the rest
Blood Glucose, Blood Glucose -- drug effects, Cardiovascular Diseases, Cardiovascular Diseases -- drug therapy, Cardiovascular Diseases -- etiology, Cardiovascular Diseases -- prevention & control, Clinical Trials as Topic, Diabetes Complications, Diabetes Complications -- prevention & control, Diabetes Mellitus, Type 2,

Named Entities:
NE form: insulin
NE type: PROTEIN
ID: Species Human (Homo sapiens): INS_HUMAN, A6XGL2_HUMAN
Species Bovine (Bos taurus): INS_BOVIN
Species Western clawed frog (Xenopus tropicalis): A4IGV9_XENTR
Species Zebrafish (Danio rerio): Q9DDE5_DANRE
Species Chicken (Gallus gallus): INS_CHICK

NE form: cardiovascular disease
NE type: DISEASE
CUI Number: C0007222

NE form: vascular disease
NE type: DISEASE
CUI Number: C0042373

NE form: type 2 diabetes
NE type: DISEASE
CUI Number: C0011860

NE form: Diabetes
NE type: DISEASE
CUI Number: C0011847, C0011849

Fig. 3 Retrieved articles marked up by semantic types

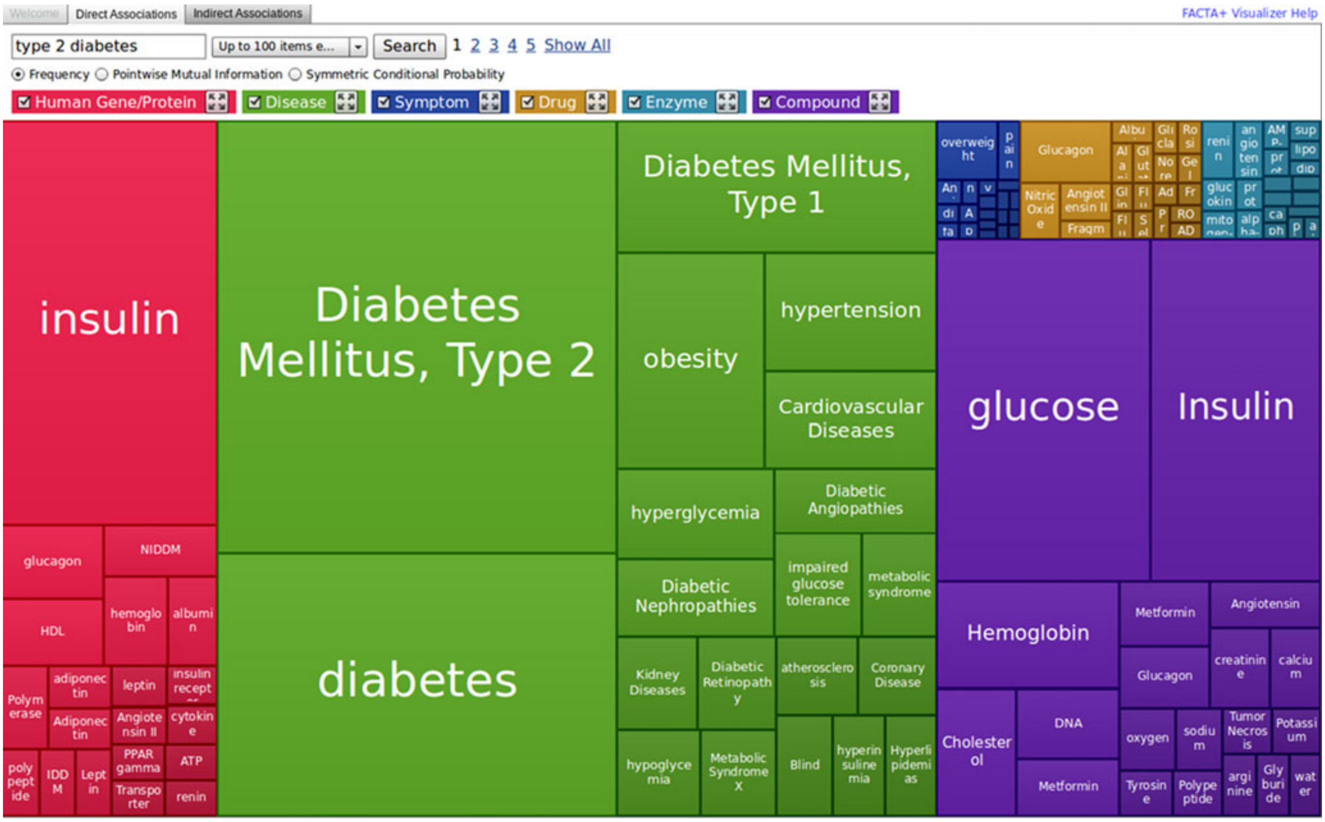
the potential association between X and Z is considered as a useful hypothesis unless there is already a known association between X and Z. This approach is often called Swanson's ABC model approach after his seminal work on literature-based hypothesis generation. Figure 4 illustrates this approach in the context of implementing it on FACTA+, where the user provides a starting query to the system 'type 2 diabetes'. We call the concepts that are directly associated with the query pivot concepts, and the concepts that are indirectly associated with the query through those pivot concepts target concepts.

Results from FACTA+, can be viewed by users in two ways: the first presents directly associated concepts (Fig. 4) and the second indirect (potentially novel) associations (Fig. 5). The concepts associated with a user query are seen as colour-coded rectangles grouped into six categories (human genes/proteins, diseases, symptoms, drugs, enzymes and chemical compounds). Initially, the number of concepts shown is limited, but more results can be visualized by applying a pager control. The importance of direct and indirect associations in relation to a query can be easily recognized by looking at the size of each rectangle representing a biomedical concept. The rectangles are arranged to maintain similar aspect ratios to make the rectangles visually recognizable. Users can also focus on a particular set of categories by using check boxes.

For directly associated concepts, each rectangle has a link to evidence sentences. When users click a rectangle, options appear to allow users to make a new search with the concept, or view a retrieved document to show 'evidence sentences' with concepts highlighted (Fig. 5).

For extracting indirect associations, pivot concepts co-occurring with the query (e.g. type 2 diabetes) are shown on the left-hand side, and target concepts co-occurring with the pivot concepts are shown on the right-hand side. When users point the mouse cursor on a particular pivot concept, visual links from the concept to its corresponding target concepts appear (Fig. 5). Similarly, when users point to a target concept, links from the concept to its corresponding pivot concepts appear.

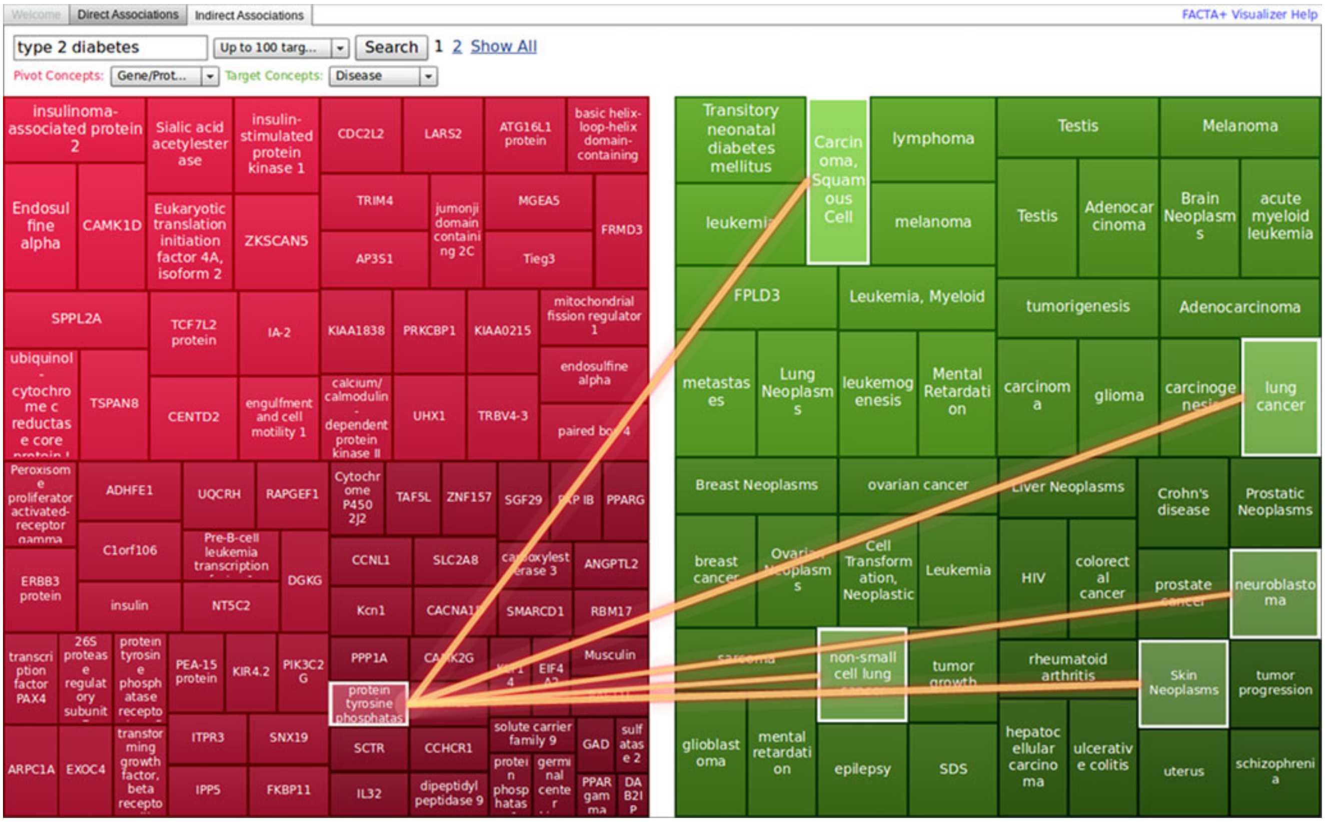
By querying with a disease name such as 'type 2 diabetes', other diseases such as 'squamous cell carcinoma', 'lung cancer', 'neuroblastoma' and 'skin neoplasms' are discovered via the pivot concept 'protein tyrosine phosphatase'. This result represents a gene/protein that is related to 'type 2 diabetes' and that is also linked to other diseases, but these



type 2 diabetes | Up to 100 items e...

● Frequency ○ Pointwise Mutual Information ○ Symmetric Conditional Probability

Fig. 4 Visualizing associations related to type 2 diabetes

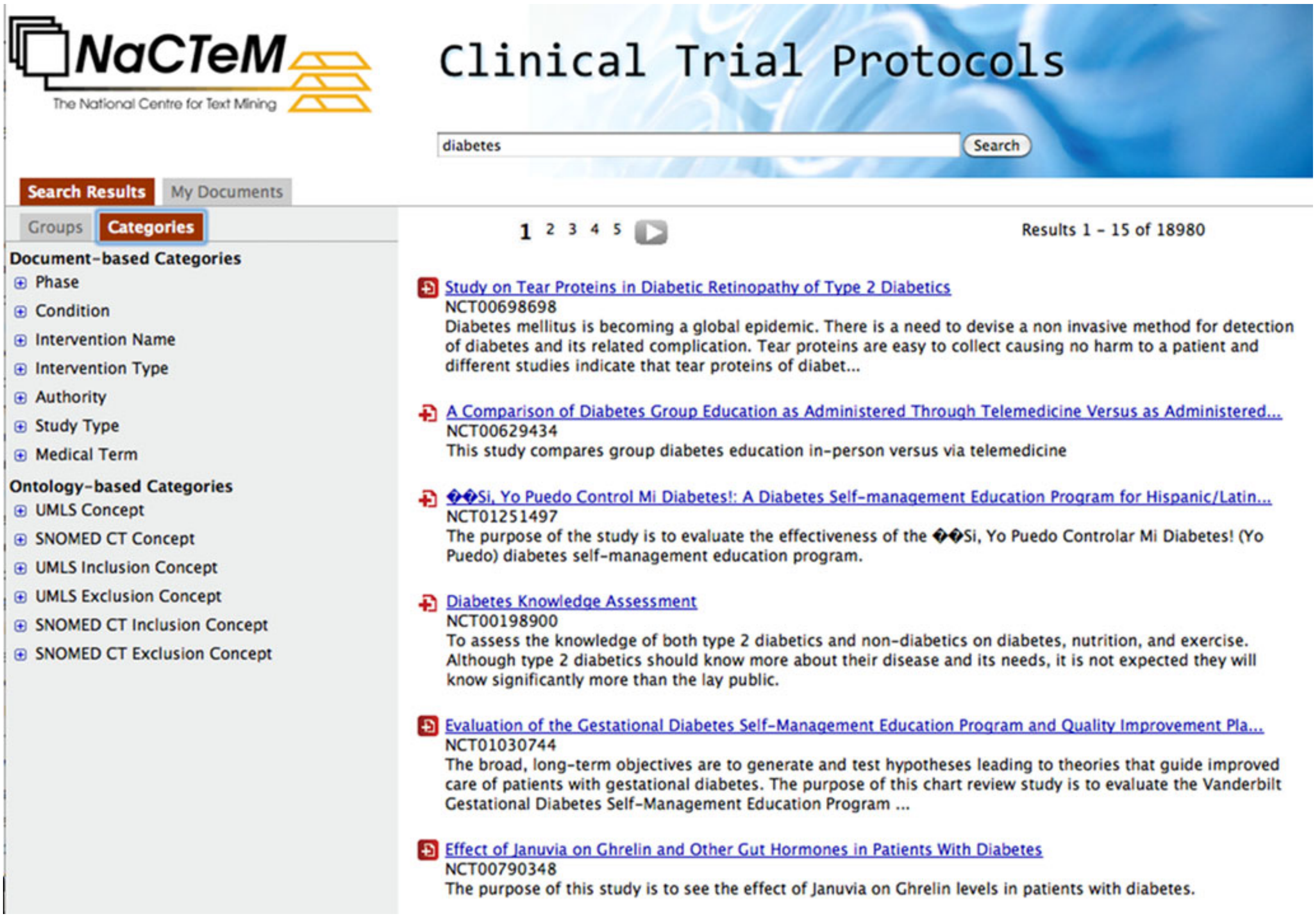


type 2 diabetes | Up to 100 targ...

Pivot Concepts: Gene/Prot...

Target Concepts: Disease

Fig. 5 Visualizing indirect associations between concepts



Study on Tear Proteins in Diabetic Retinopathy of Type 2 Diabetics
NCT00698698
Diabetes mellitus is becoming a global epidemic. There is a need to devise a non invasive method for detection of diabetes and its related complication. Tear proteins are easy to collect causing no harm to a patient and different studies indicate that tear proteins of diabet...

A Comparison of Diabetes Group Education as Administered Through Telemedicine Versus as Administered...
NCT00629434
This study compares group diabetes education in-person versus via telemedicine

��Si, Yo Puedo Control Mi Diabetes!: A Diabetes Self-management Education Program for Hispanic/Latin...
NCT01251497
The purpose of the study is to evaluate the effectiveness of the ��Si, Yo Puedo Controlar Mi Diabetes! (Yo Puedo) diabetes self-management education program.

Diabetes Knowledge Assessment
NCT00198900
To assess the knowledge of both type 2 diabetics and non-diabetics on diabetes, nutrition, and exercise. Although type 2 diabetics should know more about their disease and its needs, it is not expected they will know significantly more than the lay public.

Evaluation of the Gestational Diabetes Self-Management Education Program and Quality Improvement Pla...
NCT01030744
The broad, long-term objectives are to generate and test hypotheses leading to theories that guide improved care of patients with gestational diabetes. The purpose of this chart review study is to evaluate the Vanderbilt Gestational Diabetes Self-Management Education Program ...

Effect of Januvia on Ghrelin and Other Gut Hormones in Patients With Diabetes
NCT00790348
The purpose of this study is to see the effect of Januvia on Ghrelin levels in patients with diabetes.

Fig. 6 Semantic search based on knowledge sources

diseases do not necessarily appear together. In this way, FACTA+ potentially discovers new knowledge, hidden in the sea of literature.

EvidenceFinder, Europe PubMed Central

EvidenceFinder[9] developed for Europe PubMed Central[10] [20] is a publicly available search system based on 25 million abstracts and more than 2 million full text research articles for biomedicine. EvidenceFinder was designed for use in the context of a standard full-text retrieval service, which automatically suggests questions on the basis of what has been entered in the search engine's query field. For example, questions based on the query 'type 2 diabetes' such as 'what causes type 2 diabetes?' are automatically generated by the system. Users can then select one of the suggested questions (e.g. 'what treats type 2 diabetes mellitus', 'what prevents type 2 diabetes mellitus', etc.) and retrieve information (sentences from documents) relevant to this question from the full document collection. The approach of assisting a search strategy by generating questions matched to sentence retrieval is a promising aid to researchers.

Mining Clinical Trials

Text mining can also support a search strategy that is customized to clinical trials. One such system we have developed aims to address the information overload problem and to assist the creation of new protocols [21]. Using state-of-the-art text mining technologies, applied to large clinical trial collections, users are provided with powerful tools to narrow down their search.[11] Searches can also be conducted using semantic categories based on knowledge resources such as UMLS,[12] SNOMED CT, etc. (Fig. 6). In addition, eligibility criteria are displayed that relate to a set of chosen clinical trials to help researchers in composing new clinical trials.

[9] http://labs.europepmc.org/

[10] http://europepmc.org/

[11] http://www.nactem.ac.uk/clinical_trials/

[12] http://www.nlm.nih.gov/research/umls/

Concluding Remarks

Text mining has been used to manage the mass of literature by extracting information from text enabling researchers to discover, collect, interpret, synthesize, select and organize (curate) knowledge. Text mining techniques can be applied in a variety of areas of medicine and can include text types such as full papers, abstracts, clinical trials and even electronic health records.

Semantic text mining techniques can be customized to extract semantic types, relations and associations with multifactorial diseases such as diabetes. Currently, such extraction is being manually conducted by a large group of scientists, and therefore it is anticipated that text mining will contribute to the automation of this work.

In this contribution, we have illustrated the potential benefits of text mining approaches by using the example of diabetes because it is an important and complex disease that has multiple aetiologies [22]. Researchers who are currently struggling to cope with the large amount of complex literature on diabetes could benefit particularly from the powerful capabilities offered by text mining.

We predict that new technologies such as text mining will have a positive impact on diabetes research and research into other complex diseases. This should lead to the more efficient use of resources, better quality research and ultimately to improved disease prevention and therapy.

References

1. Ananiadou S, McNaught J, editors. Text mining for biology and biomedicine. London: Artech House; 2006. p. 286.
2. Ananiadou S, Kell DB, Tsujii J. Text mining and its potential applications in systems biology. Trends Biotechnol. 2006;24(12):571–9.
3. Hunter L, Cohen KB. Biomedical language processing: what's beyond PubMed? Mol Cell. 2006;21:589–94.
4. Zweigenbaum P, Demner-Fushman D, Yu H, Coehn KB. Frontiers of biomedical text mining: current progress. Brief Bioinform. 2007;8(5):358–75.
5. Cohen KB, Hunter L. Getting started in text mining. PLoS Comput Biol. 2008;4(1):e20.
6. Rzhetsky A, Seringhaus M, Gerstein M. Seeking a new biology through text mining. Cell. 2008;134(1):9–13.
7. Frijters R, van Vugt M, Smeets R, van Schaik R, de Vlieg J, Alkema W. Literature mining for the discovery of hidden connections between drugs, genes and diseases. PLoS Comput Biol. 2010;6(9):e1000943.
8. Tsuruoka Y, Miwa M, Hamamoto K, Tsujii J, Ananiadou S. Discovering and visualizing indirect associations between biomedical concepts. Bioinformatics. 2011;27(13):i111–9.
9. Okazaki N, Ananiadou S. Building an abbreviation dictionary using a term recognition approach. Bioinformatics. 2006;22 (24):3089–95.
10. Okazaki N, Ananiadou S, Tsujii J. Building a high-quality sense inventory for improved abbreviation disambiguation. Bioinformatics. 2010;26(9):1246–53.
11. Miyao Y, Ohta T, Masuda K, Tsuruoka Y, Yoshida K, Ninomiya T, Tsujii J. Semantic retrieval for the accurate identification of relational concepts in massive textbases. Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics; 2006. Stroudsburg, PA: Association for Computational Linguistics; 2006. p. 1017–24.
12. Pyysalo S, Ohta T, Miwa M, Cho HC, Tsujii J, Ananiadou S. Event extraction across multiple levels of biological organization. Bioinformatics. 2012;28(18):I575–81.
13. Wang X, McKendrick I, Barrett I, Dix I, French T, Tsujii J, et al. Automatic extraction of angiogenesis bioprocess from text. Bioinformatics. 2011;27(19):2730–7.
14. Ananiadou S, Pyysalo S, Tsujii J, Kell DB. Event extraction for systems biology by text mining the literature. Trends Biotechnol. 2010;28(7):381–90.
15. Ananiadou S, Nenadic G. Automatic terminology management in biomedicine. In: Ananiadou S, McNaught J, editors. Text mining for biology and biomedicine. London: Artech House; 2006. p. 67–97.
16. Tsuruoka Y, McNaught J, Ananiadou S. Normalizing biomedical terms by minimizing ambiguity and variability. BMC Bioinforma. 2008;9 Suppl 3:S2.
17. Weeber M, Vos R, Klein H, De Jong-Van Den Berg LT, Aronson AR, Molema G. Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide. J Am Med Inform Assoc. 2003;10(3):252–9.
18. Swanson D. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspect Biol Med. 1986;30(1):7–18.
19. Swanson D, Smalheiser N. Assessing a gap in the biomedical literature: magnesium deficiency and neurologic disease. Neurosci Res Commun. 1994;15:1–9.
20. McEntyre JR, Ananiadou S, Andrews S, Black WJ, Boulderstone R, Buttery P, et al. UKPMC: a full text article resource for the life sciences. Nucleic Acids Res. 2011;39(Database issue):D58–65.
21. Korkontzelos I, Mu T, Ananiadou S. ASCOT: a text mining-based web-service for efficient search and assisted creation of clinical trials. BMC Med Inform Decis Mak. 2012;12 Suppl 1: S3.
22. Yach D, Stuckler D, Brownell KD. Epidemiologic and economic consequences of the global epidemics of obesity and diabetes. Nat Med. 2006;12(1):62–6.