Literature-Based Discovery: Critical Analysis and Future Directions

Ali Ahmed

Cairo University, Egypt

Summary
Literature-Based Discovery (LBD) is the science of relating existing knowledge in literature to discover new relationships. It is sometimes referred to as hidden knowledge. The paper provides the most recent classification of the existing LBD methods relating the problem to other domains such as information retrieval. The paper identifies that Vector Space Model, Probabilistic Model, and Inference Network Model are the mostly used for LBD problem. The researchers of this paper justified their belief that there are important differences between the two problem domains with regards to novelty, time factor, reasoning, and relevance. The paper investigates the hypothesis that some discoveries could have been materialised earlier based on some early relatedness indicators. The latter point is an interesting one that offers some direction for the future research in LBD. Moreover, the paper introduces the on-going work of the author on proposing a new evaluation methodology that addresses the weaknesses of the current methodologies investigating the desirable characteristics of the future LBD evaluation methodology.

Key words:
Digital Privacy, Island of Jersey, jurisdictions, Employee Rights.

1. Introduction

Discovery in science is the result of the formulation of novel, interesting, and scientifically sensible hypotheses. These hypotheses can be formulated by reviewing the existing body of domain-specific literature. The voluminous amount of data stored in the literature, however, makes the task impossible to be performed manually by scientists. Literature-based discovery (LBD) is a type of text mining that aims at identifying nontrivial assertions that are implicit within a large body of documents [1]. LBD holds the potential to help scientists particularly in biomedicine and genomic to accelerate their scientific discovery progress by automating the generation of viable scientific hypotheses. To achieve such a purpose, there is a fundamental need for classifying the current work in the field relating the respective works to their area of research to draw the roadmap of the research in this hot point. Thus, this work provides the most recent classification of the existing LBD methods relating the problem to other domains such as information retrieval. It also draws the countries between the two main problem domains of information retrieval and LBD. The paper also investigates that some literature discoveries could have been materialised earlier based on some early relatedness indicators. In addition, the paper introduces the on-going work of the author on proposing a new evaluation methodology that addresses the weaknesses of the current methodologies investigating the desirable characteristics of the future LBD evaluation methodology.

This work comprehensively reviews the academic literature as well as Pubmed [a] to classify the LBD proposals in order to compare them, to investigate the existence of early relatedness indicators for literature-based discoveries, and to assess the current LBD Evaluation methodologies against a set of gold standards that defines what an evaluation methodology should be. The following research questions are to be answered for such a purpose:

1. What are the current categories of LBD methods? This research question tries to find relationships among various LBD proposals in the academic literature. In addition, what is the prevalent category in LBD? Identification of such a category helps to study its characteristics and whether it seamlessly meets the characteristics of the problem in hand. For example, what are the important differences between Information Retrieval domain and LBD domain? In order to answer those research questions, the survey protocol will incorporate the research works with high impact as well as citation index. Works that have not yet been cited or works that replicate current results will not be surveyed in this study.
2. Could some discoveries have been materialised earlier based on some early relatedness indicators? In order to answer this important research questions, PubMed will be used also as a source of accredited research works and to investigate the possibility of early relatedness indicators discovery.

[a] The US National Library of Medicine National Institutes of Health, http://www.ncbi.nlm.nih.gov/pubmed

Manuscript received July 5, 2016
Manuscript revised July 20, 2016

3. What are the problems with the current LBD evaluation methodologies? For instance, do the differences between various categories of LBD affect the quality of the proposals? For example, the IR-centric evaluation methodologies (i.e. those relied on ranking metrics and Precision-Recall scheme) could not be utilised for other LBD categories. In other words, given the same measures to be compared, will their evaluation yield similar result if they are evaluated on a different platform? Is there any standard for evaluation methodologies? How about gold standards?

The organisation of this papers follows the following:

1. Section 2 studies the current LBD research proposal to classify the different LBD proposal. Based on the literature survey, the paper identifies that Vector Space Model, Probabilistic Model, and Inference Network Model are the mostly used for LBD problem.
2. Section 3 distinguishes between OR and LBD problem domains.
3. Section 4 discusses the relatedness indicators for early discovery in literature. The section suggests that indirect connections between concepts can be predicted much earlier by looking at interesting patterns and changes over the citation network space. It is possible to think that the sudden emergence of publications concerning concepts at roughly the same time could have been a natural response to a specific, significant event (e.g. a prior scientific discovery, a discovery of a new drug, etc.).
4. Section 5 studies how to evaluate an LBD proposal and proposes a gold standard for such a purpose. The implication of differences in the current proposal, as well as the well-perceived methodologically-flawed, from the perspective of the author, proposal of Yetisgen-Yildiz and Pratt [2] gives us the incentive to re-evaluate and criticize the current evaluation methodologies which have heavily relied on ranking metrics and Precision-Recall scheme (IR-centric).
5. Section 6 concludes the work and highlights the future directions of this research.

2. LBD: Literature Survey

Several LBD methods have been proposed which focus on the analysis of scientific documents such as journal articles. In fact, the field started with a trial-and-error model that studies two groups, A and C, in a bibliographic collection in terms of their common descriptors (i.e. indexing terms), mutual citation, bibliographic coupling, and co-citation. This is mainly Swanson work in 1987[3]. The model manually studies A and C groups exhaustively. It is almost impossible to apply the model on a large corpus of documents. Statistics and probability could be used as a way to identify new discoveries. Probability is utilised in Information Retrieval (IR) as well as LBD. Singhal work in [4] falls under the former category where ranking is used. In fact, the work is based on the general principle that ranks documents in a particular collection by decreasing probability of their relevance to a certain query.

In general, models in this category inherently entail different forms of ranking mechanism. LBD Utilises also various statistical and probabilistic relevance measures to infer indirect relationships between A and C (i.e. the Swanson's ABC model). The models under this umbrella address the Open Discovery Problem (ODP). Given a starting A term, select and rank a list of B terms with high probability of being relevant to A. For each selected B term, find C terms highly relevant to each B term. If A does not overlap with C, an indirect relationship between them is probable. It is often assumed that the more intermediate B terms shared by a pair of A and C, the stronger their indirect relationship would be. Most of the models under this category rely on the term co-occurrence without much reliance on domain-specific knowledge sources. Thus, the use of probabilistic methods can be easily extended to other domains. Table 1 summarises the proposed statistical/probabilistic methods.

Table 1. Statistical Methods used in LBD

Annotated Description	**Proposal(s)**
Lexical statistics, TF*IDF	[5,6,7]
Combines strength of direct associations and reliability of indirect association of concepts	[8,9,10]
Statistical analysis of gene-disease occurrences in the biomedical literature	[11]
Association rules, Term-frequency scheme, Use citation information, Non-binary term weighting	[12,13,14,15]
Mutual Information Measure, R-score	[16,17,18]
BioBibliometric Distance, Dice coefficient, Visualization of gene network	[19]
Entity-based network, Minimum Mutual Information Measure (MMIM)	[20,21,22]

Vector Space Model (VSM) / Algebraic in IR, a document and a query are represented by a vector of terms. A document's score is given based on measuring the similarity between the query (i.e. query vector) and the document (i.e. document vector). Cosine and Inner-product between two vectors are commonly used as the numeric similarity [4]. In LBD, the approach establishes the ABC model based on document similarity even though A and C terms do not co-occur. This is a stark difference to the statistical methods discussed earlier. The proposals in this category are generally marked by their utilisation of (i) vector representation and vector algebra, (ii) document similarity measures, (iii) term by document matrices, and (iv) representations of terms and documents within hyper-dimensional spaces. Table 2 summarises such proposals.

Table 2. VSM used in LBD

Annotated Description	Proposal(s)
Abductive reasoning, Reflective Random Indexing, Distributional semantics, Quantitative estimation of term similarity, Semantic space, Dimensionality reduction	[23,24,25,26,27, 28,29,30,31,32]
Latent Semantic Indexing (LSI), Singular Value Decomposition (SVD), Ranking by Cosine Similarity	[33]
Stepping Stones and Pathways (SSP), Document similarity, Bayesian Network, Citation analysis	[34,35]
Vector of sub-vectors,Weighted term vectors by TF*IDF, Topic profile, Cosine Similarity, Semantic-type filtering	[36,37,38,39,40, 41,42]
Context term vector, Cosine Similarity, Spearman Correlation	[43,44,45]
Weighted concept fingerprint/profile, Proximity of concepts in vector space, Similarity of concept fingerprints, Cluster analysis, Path-finding	[46,47,48,49,50, 51,52,53,54,55,56, 57]
Semantic features, Dimensionality reduction, Gene-document matrix, Clustering	[58], [59]
Feature vectors, Cosine Similarity, Clustering	[60]
Outlier detection, Similarity graph, Ensemble heuristics	[61,62,63,64,65, 66,67,68]
Conceptual network, Lnu Weighting	[69], [70]
Compound correlation model, Cosine Similarity	[71]
VSM extended with Transitive Closure, Combine VSM with inference Process	[72]
Abductive reasoning, Quantum Informatics, Information Flow	[73,74,75,76,77, 78,79,80,81,82,83, 84,85]
Latent Semantic Indexing (LSI), Implicit gene relationships, Identification of transcription factor candidates, Nonnegative Matrix Factorization (NMF)	[86,87,88]
Matrix decomposition, Factor screening, Eigenvector, Transitive text mining	[89]

Knowledge-based methods have been used in a three-fold. Firstly, in Artificial Intelligence (AI), the approach is characterized by a particular focus on the accumulation, representation, and use of knowledge specific to a particular task. The source of the system's power is the task-specific knowledge rather than domain-independent methods. Two components central to the operation of such system are the knowledge base and the inference engine. Secondly, Knowledge-Based IR employs rich knowledge representations. Two predominant approaches have been used to develop IR systems where knowledge-based intelligence resides in [90] (a) the interface to a traditional IR system; and (b) the representational formalism of the information stored in the IR system. Knowledge could come different forms such as frames, semantic nets, production rules, etc. Thirdly, in LBD, the approach is characterized by heavy reliance on domain-specific knowledge sources (e.g. ontologies, knowledge bases, inference rules). As a result, it is typically difficult to extend the approach to other domains. the approach could be categorized further according to the specific technique being used into:

- Semantic filtering
- Subsumption reasoning based on ontology
- Semantic similarity
- Association and annotation detection
- Biological network/graph and path analysis
- Rule-based reasoning
- Cluster analysis

Table 3 summarises the proposals under this category.

Table 3. Knowledge-based Methods used in LBD

Annotated Description	Proposal(s)
Concept-based, Log-likelihood ratio,Word-frequency ranking, Semantic type filtering	[91, 92, 93, 94]
Similarity in annotated phenotypes in ontologies, Ontology subsumption reasoning	[95]
Semantic similarities between events, Information Content (IC)	[96,97]
Semantic similarity, Association score computed using a regularized Log-Odds score, Resnik Similarity	[98,99,100]
Chemical, diseases, proteins, Proteins as B-terms	[101]
Gene expression profiles	[102]
Ontology, Subsumption, transitivity, and domain-oriented rules	[103]
Interaction Network, Network centrality measures (degree,	[104,105,106, 107,108]

eigenvector, betweenness, and closeness measures), NLP	
Biomedical concept network, Neighborhood measures, Number of paths, Distance	[109,110,111]
Biological Distance, Discrimination of gene pathways	[112,113,114]
Discover diseases that are connected to the same pathways, Network Analysis	[115]
Biological network, Qualities: relevance, informativeness, and reliability, Proximity measures	[116]
Semantic predication, Network analysis, Degree of centrality	[117,118]
Automated reasoning, Logic Rules, Logic Facts, NLP	[119,120,121,122, 123,124]
Association Profile, Cluster Analysis, Regularized Log-Odds Function, Term statistical distribution	[125,126,127, 128]
Cluster analysis, Instance-based learning	[129]

The Inference Network basically deals with IR as well as LBD. In the former, document retrieval is conceived as an inference process in an inference network [4]. According to Turtle and Croft, the basic document retrieval inference network consists of a document network and a query network [130]. The former component decomposed of document nodes, text representation nodes and concept representation nodes. A document node represents a document in the collection. In fact, the query network is modelled as an "inverted" acyclic dependency graph (ADG). The ADG of the query network has a single node (i.e. leaf) that corresponds to the event that an information need is satisfied. It has also multiple nodes (i.e. roots) that express the information needs. The significant probabilistic dependencies are captured by the retrieval inference network. In fact, those dependencies represent the significant probability amongst the variables represented by the nodes in both document and query networks. The node belief is computed once the build of the query network is done. The initial value at the information need node is the probability that the information need is met given no particular document has been observed as well as all documents are equally likely or unlikely. On the other hand, the work done by Seki [131] is considered LBD. Basically, to model gene-disease associations, a disease is treated as a query node and genes as document nodes. Connecting these nodes exhibits two types of intermediate nodes: gene functions and phenotype nodes which characterize the genes and disease, respectively. The edges between these nodes are established based on the existing knowledge stored in knowledge bases and literature. After constructing the inference network, causative gene set G for given disease d is predicted by probability measures. In fact, compared to the VSM, this model has the advantage of incorporating multiple intermediate nodes [131]. To summarise, the work uses an extended inference network as well as ontology to enhance probability estimates which is actually utilises the conditional probability [132].

Intellectual Structure Analysis technique is used in Scientometrics. Based on Chen [133], the main goal is to identify what kind of information could be considered as early signs of potential discoveries. The Structural Variation approach is centred on the novel boundary-spanning connections introduced by new articles. The theoretical foundation is simply that boundary-spanning, brokerage, and synthesis mechanisms in an intellectual structure can explain the scientific discoveries. The change in the structure based on the introduction of a new article is measured by the Cluster Linkage (CL). The change is actually tangible in terms of new connections added between clusters. CL was found to be the strongest predictor for an increase in citation counts. Adopting the Intellectual Structure Analysis in LBD requires a representation of the intellectual structure. The intellectual structure could be formed differently based on co-citation either for references or authors, or co-occurring keywords. Chen's method is a generalized form of Swanson's ABC model. To connect A and C, it does not require existing relationships through the ABC path. It is also not limited to three entities. It addresses the novelty of a connection that links groups of entities as well as connections linking individual entities [134, 135,133].

The Fuzzy sets theory can deal with this kind of a problem. In IR, the concept of document relevance to a particular keyword query follows, actually, a fuzzy logic-based interpretation. A logical model of IR was developed that accounts for imprecise and uncertain information via the use of fuzzy logic, which: (a) assumes linguistic terms as importance weights of keywords in documents; (b) considers the uncertainty of documents and queries representation; (c) interpret the linguistic terms in the representation of documents and queries as well as their matching in terms of the Zadeh's fuzzy logic [136]. The Fuzzy Sets theory is applicable in LBD domain. Based on Wren's interpretation of the ABC model, "the fuzzy set theory replaces the two-valued set-membership function with a real-valued function. Membership of C in A is treated as a probability or as a degree of relatedness. When asserting a relationship, a real value is assigned to assertions as an indication of their degree of relatedness, which ranges from 0 (unrelated) to 1 (identity) as shown in Figure 1. Fuzzy set membership is shown by sub-figure (b). The domain of a given term is defined by the relative frequencies of all the other terms it is co-occurred with in the literature. The overlap that a term has with any other term is a function of the terms they co-occur with and the relative importance of this shared term to both domains. The result is a method able to identify highly similar biomedical concepts and properties" [137]. Two different fuzzy binary relations are defined, one between disease and drug terms, the other one between drug and gene ontology terms [138]. It is assumed that two terms are highly related if they appear frequently together. The strength of association is estimated by counting the co-occurrences of both terms in the same 'transaction' (i.e. literature abstracts).

Some web data mining proposals are related to the problem in hand. The goal is to find a good path between two articles. The path is referred to as a story between the articles. An emphasis is placed on forming a coherent chain [139]. Kumar is calling it storytelling formulating the problem as a generalization of redescription mining [140]. Storytelling aims to explicitly relate object sets that are disjoint by finding a chain of approximate re-descriptions between the sets. The strength of the story is determined by the weakest transition. In LBD, the most salient example of the application of storytelling algorithm is given by Hossain [141]. "Given a start and an end publication (with little or no overlap in content), it identifies a chain of intermediate publications from

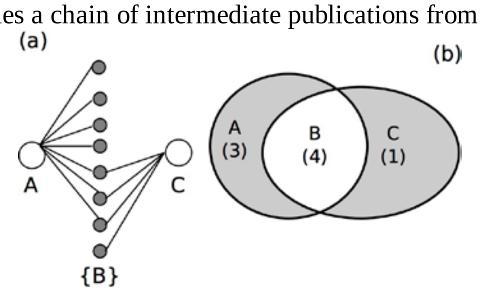

Fig. 1. Wren's interpretation of the ABC model [137]

one to the other such that neighbouring publications have significant content similarity" [142]. Despite its utilization of some forms of similarity measures (i.e. the primary focus of VSM), this method is distinct in that: (i) it involves longer chain of documents; (ii) its particular emphasis on building coherent and biologically interpretable 'story'; and (iii) it does not materialize a complete similarity graph which is computationally expensive. Proposals such as [143, 144, 145, 142, 141, 139] fall under such a category. Tools such as Generalization of Re-description Mining, Cohesiveness, Path-finding, Weighted term vector, Soergel Distance, Nave Bayes Classifier, and NLP are utilised by those proposals.

Database Tomography (DT) exhibits some resemblance to the Cluster-based Retrieval used in in IR. Cluster-based retrieval is based on the hypothesis that similar documents will match the same information needs. The method groups documents into clusters and return a list of documents based on the clusters that they come from. One approach "is to retrieve one or more clusters in their entirety in response to a query. The task for the retrieval system is to match the query against clusters of documents instead of individual documents. It then ranks clusters based on their similarity to the query. The second approach to cluster-based retrieval is to use clusters as a form of document smoothing. Previous studies have suggested that by grouping documents into clusters, differences between representations of individual documents are, in effect, smoothed out. Cluster-based Language models have been employed in topic detection and tracking. Document clustering is used to organise collections around topics. Each cluster is assumed to be representative of a topic, and only contains stories related to that topic" [146]. A cluster-based retrieval using language model builds a language model for each document in the collection, and rank the documents according to the probabilities that a query could have been generated from each of these document models. DT was introduced by Kostoff as "a revolutionary approach for identifying pervasive themes and thrust areas intrinsic to textual databases, the connectivity among these areas, and sub-thrust areas closely related to and supportive, of the pervasive thrust areas" [147]. Adapting DT to IR, it resulted in a method called Simulated Nucleation, in which a small core group of documents relevant to the topic of interest is first retrieved. Next, patterns of word combinations in existing fields are identified, new query term combinations based on the newly-identified patterns are generated, and the process of retrieval is repeated. In addition, patterns of word combinations which reflect extraneous non-relevant material are identified, and search terms which have the ability to remove non-relevant documents from the database are inserted. The nucleus continually expands its coverage and improves the quality of the core. This iterative procedure continues until convergence is achieved where relatively few new documents are found even though new search terms are added. DT operates on top of word frequency and word proximity analysis. Simulated Nucleation organizes documents into theme-oriented clusters similar to cluster-based retrieval. Its emphasis on topic/theme detection renders some similarity to the goal of cluster-based language retrieval models.

Kostoff 's work stream is interesting indeed [147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165]. The following summarises the approach:

1. Retrieve core literature to target problem (C).

   - Generate query for core literature
   - Enter query into database search engine and retrieve core literature
2. Characterize core literature.
   - Obtain technical infrastructure (people, institutions) of core literature through bibliometrics
   - Obtain technical structure of core literature (themes, relationships among themes) through NLP. Cluster core literature records to identify key technical thrusts.
3. Expand core literature.
   - Generalize query term for each key technical thrust identified above
   - Retrieve literature related directly and indirectly to each key technical thrust
4. Generate potential discovery.
   - Restrict classes of solutions based on semantic categories
   - Examine all remaining records
   - For records that appear to contain potential discovery, perform vetting procedure to ensure genuine discovery

To ensure the completeness of the retrieved core literature at the initial phase, the author makes use of long query statement consisting of up to hundreds of keywords, co-occurrence phenomena, and latent feature indexing. The combination of cluster formation, query expansion, iterative retrieval and relevance feedback makes this approach unique from the other methods.

There is another approach that is a hybrid of Statistical/Probabilistic and Knowledge-Based. Methods in this category do not fit into the pure statistical/probabilistic model due to the significant role (particularly in the form of semantic-based filtering) that domain-specific knowledge sources (e.g. ontologies) play in increasing the systems' precision. On the other hand, the ability of the systems to make inferences relies on a range of statistical methods. The latter studies in IR showed the possibility of integrating statistical and knowledge-based IR methods. Table 4 summarises the work done under this category in LBD.

Table 4. Statistical Methods used in LBD

Annotated Description	Proposal(s)
Z-score, Association rules, Ranking based on the number of linking terms between the starting and target terms	[166,167,168]
Association rules, Ranking by Confidence value, Semantic type filtering, Ranking by number of intermediate paths between A and C, Semantic Predications	[169,170,171,172, 173,174,175,176]
Identifies interacting patches of proteins based on hydrophobicity, accessibility and residue interface propensity, Atomic distance	[177]
Association rules, Semantic type-filtering, Ranking by B-term count, Ranking by F-measure	[178,179,180,181, 129]
Statistical criteria, Term-frequency probability cut-off, Semantic-type Filtering	[3,182,183,184, 185,186,187,188, 189,190]
Association rules, Concept siblings, Ontology, Concept replacement	[191,192,193, 194]

Density-based Clustering method organizes clusters as dense regions of objects that are surrounded by regions of low density. This cluster definition is often employed when the clusters are irregular or intertwined, and when noise and outliers are present. Cluster analysis aims to use a variety of cluster properties as predictors of interesting patterns. Consequently, although DT utilizes clusters as part of its methodology, it is distinct from cluster analysis (in data mining sense) in that clusters are used in DT only for organizing documents around specific themes rather than for prediction purpose. Stegmann highlighted that to replicate Swanson's fish oil-Raynaud's disease hypothesis discovery, a set of Raynaud's disease documents were downloaded and important terms were extracted [195, 196]. Next, high Equivalence Index term pairs were clustered. Cluster properties (density, centrality) were computed. Maps of density and centrality were generated. Examination of the map revealed interesting term pairs at the lower left quadrant (including the intermediate B-terms and fish oil term).

3. LBD Vs IR Problem Domains

From the discussion of existing proposals, one can observe that Vector Space Model, Probabilistic Model, and Inference Network Model are the mostly used. Gordon et al also distinguished LBD from knowledge discovery and data mining in that LBD seeks for relationships that may exist beyond a defined set of texts [6]. Beyond this point, the relationship between LBD and data mining has not been made much clearer. Data mining methods (cluster analysis, storytelling) are seen in storytelling and cluster analysis. Unsurprisingly, knowledge-based approach characterizes many existing methods (i.e. knowledge-based) because of LBD's domain-specific nature and its demand for substantial logical capability in order to increase the precision of its results. Likewise, the hybrid use of the probabilistic and knowledge-based models is necessary to balance the trade-off between recall and precision. The rest of categories are filled by unique approaches (Database Tomography, Fuzzy Set Theory and Intellectual Structure Analysis) that have their bearing on the solution to LBD problems. The dominance of IR-based models suggests that LBD is seen as a sub-specialization of IR problem, except that LBD address a much harder problem [6]. However, we believe that there are important differences between the two problem domains with regards to novelty, time factor, reasoning, and relevance. For instance, the time-line is an interesting factor in LBD literature. The time factor may have a significant bearing on the mechanism of scientific discoveries in general. The questions here are simple: Could Swanson have formulated his hypothesis much earlier than 1986? How early can the hypothesis be actually made? It seems plausible to assume that it is important for the bodies of literature for Fish Oil, Blood Viscosity, and Raynaud's Syndrome to grow and reach their "critical mass" such that the inferred relationship between them can be possibly hypothesized. But when is this critical mass achieved? Could it have happened much earlier than 1986? How is it measured? These are important questions for which we don't have the answer yet (to the knowledge of the author). For instance, this has an implication on the evaluation methodology for LBD. Until now, it seems safe to claim that there is no proper evaluation methodology for LBD. Without it, there is no good way to compare the performance of the existing systems. This is a quite interesting topic that will be discussed further later on in the early relatedness indicators section. LBD relies on IR-based "techniques and insights, but is a much harder problem. Whereas IR has, at the outset, the objective of finding documents relevant to a given need for information, the success of literature-based retrieval depends on finding topics (or documents) that are only indirectly relevant to the topic one uses to initiate the discovery process. In addition, what is found must be previously unknown in relation to the starting point." [6]. Kostoff supports such an argument by stating, we believe there is no scientific basis for such ranking metrics and their use militates against the more infrequent concepts that could represent radical discovery" [197].

4. Early Relatedness Indicators

Is there a common phenomenon between the related concepts in publication? For instance, does the number of publications of those related concepts increase coincidently in a certain time period? Could this increase be seen as a potential relatedness? Such an increase in the number of publications could be a response to the same stimulus (i.e a real-world event). But these connections cannot be directly observed over the citation network space [133], for instance, because they are quite distantly separated. In other words, borrowing Swansons terminology, they are completely disjointed. Thus, the question is simply, "In other words, could some researchers have noticed the connection between fish oil and blood viscosity much earlier than the publication in 1984?" Wren highlighted, "One possible way of addressing this might be to turn to historical analysis. If historical relationship networks could be created, we could study how they have evolved over time, asking the critical question: How many scientific discoveries known today would have been highly ranked inferences in the past based solely upon what was known at the time? More specifically it can be asked how well any particular approach would have performed historically in predicting the probability an implicit relationship will of future scientific relevance." [198]. To investigate that hypothesis, Pubmed is consulted for the following:

– Searching for "Fish Oil". The earliest paper was published in 1926 followed by the next publication in 1945. There is no publication in the years between. From 1945 onwards, the number of publication started to increase.
– Searching for "Raynaud Disease". The earliest publication appeared in 1945 (i.e. 4 papers) from which the frequency increased quite obviously.
– Searching for Pubmed for "blood Viscosity" and found that the earliest papers were published in 1919 (2 articles) and 1927 (1 paper). The next publication, interestingly, only came in 1945 (2 papers) followed by a steady number of publications in the following years.
– Searching for "Blood Viscosity" AND "Fish Oil" and found that the earliest paper was published in 1984, only two years before Swanson published his hypothesis in 1986.
– Searching again for "Raynaud Disease" AND "Blood Viscosity" and found that the earlier paper was published in 1965.
– Lastly, to ensure that these findings are not biased by Pubmeds inherent limitation, we searched Pubmed for one of the oldest-known disease "Cholera" and found that Pubmed was able to track publication as far back as 1821.

The frequency of publications for all three concepts (individually) showed very similar patterns, pointing to a common year (i.e. 1945) from which the frequency of publication subsequently increased. Why year 1945? this is beyond the scope of this research. But searching Pubmed again for a random disease (in this case for pneumonia), no similar pattern was found. Year 1945 does not appear to be a significant year in the case of pneumonia though. Although the earliest publication for “Blood Viscosity” AND “Fish Oil” only appeared in 1984, is it possible that an inference about their connection be made much earlier but unpublished? The results of a Pubmed search is a function of the all the keywords used in searching by the user, in indexing by MEDLINE indexers, and in expressing thoughts and ideas by the authors. What if fish oil was more well-known by another term, e.g. “Fatty Acids, Omega-3” in the past? Or, what if the author chose to use a term other than “blood viscosity” to refer to the same concept? For instance, when adopting more generalized search keywords (“Fish Oil” AND “Blood”), the earliest paper appeared in 1946 entitled ‘Survival time of hypertensive rats receiving fish-oil extracts’. Interestingly, blood viscosity is a factor affecting arterial blood pressure (i.e. hypertension). To further prove this, a subsequent search for “Blood Viscosity” AND “Arterial Tension” revealed that the earliest article has appeared in 1958 (26 years before 1984!). Therefore, it is plausible to argue that some researchers might have noticed the connection between fish oil and blood viscosity much earlier than the publication in 1984. It appears that indirect connections between concepts can be predicted much earlier by looking at interesting patterns and changes over the citation network space. It is possible to think that the sudden emergence of publications concerning the three concepts at roughly the same time (i.e. 1945) could have been a natural response to a specific, significant event (e.g. a prior scientific discovery, a discovery of a new drug, etc.). This common response may serve as a very early sign of their relatedness (which became obvious only decades later). However, this requires further investigation.

5. Evaluation Methodology and Gold Standards

The implication of those differences gives us the incentive to re-evaluate and criticise the current evaluation methodologies which have heavily relied on ranking metrics and Precision-Recall scheme (IR-centric). To the best knowledge of the author, the best evaluation methodology to date is Yetisgen-Yildiz and Pratt [2]. But the paper has a methodological flaw: the authors used their own LBD systems, called LitLinker, as the platform on which the effectiveness of four (4) correlation measures were compared. This means that the evaluation process is biased towards LitLinker’s technical features (e.g. it represents the content of MEDLINE documents using the MeSH index terms which is not necessarily the best form of representation). While it makes perfect sense to compare the correlation measures against the same baseline mechanism (i.e. LitLinker), we don’t know to what extent that LitLinker’s technical biases have affected the discovery power of each measure. In other words: given the same four measures to be compared, will their evaluation yield similar result if they are evaluated on a different platform other than LitLinker? No one is sure of the answer.

In our opinion: (a) a good evaluation methodology should not be implemented upon a specific system in order to avoid biases; (b) LBD systems should be evaluated at the systemic level, not just by comparing the specific measures/algorithms implemented by the systems. For instance, it is quite obvious that the way the documents are represented (the input format) will affect the effectiveness of the discovery. Discovery outcome will differ between those who use title only and those who use full-text. Consequently, it sounds possible, given:

- A target discovery D_t
- A collection of literature L before the publication of Dt
- LBD systems to be evaluated (S_1 , S_2 , , S_n) The best performing system should:
    - Successfully draw a hypothesis concerning Dt
    - Brings D_t to the attention of the user requiring minimum amount of user’s cognitive load. For instance, if a ranking mechanism is employed, Dt should be ranked highly.
- Detect D_t as the earliest point over the publication time-line based on literature set L. This is where time factor plays a crucial role in discovery process. LBD systems should be measured based on their ‘insightfulness’. An analogy is suitable here: at the same point in time and given the same access to information, an insightful field expert is more likely to be able draw future new correlations, relevance, or possible discovery concerning a particular scientific field in comparison to a fresh PhD graduate from the same field. We say that the expert has greater insight into the field than the fresh graduate. Similarly, it is plausible to assume that a more ‘insightful’ LBD system S is able to reach D_t with less amount of information from L compared to other less insightful LBD

systems. In other words, a better LBD system is able to discover D_t at much earlier time.

Kostoff highlighted, “A central problem with all the LBD studies that have been reported in the open literature is the absence of a gold standard that can be used as a basis of comparison” [198]. Wren also noted, “Currently, it is not at all clear which LBD approaches are most efficient due to a lack of quantitative methods and gold standard test sets for analysis” [198]. Although Yetisgen-Yildiz and Pratt identified four current evaluation approaches, those categories are actually falling into two broad ones (i.e. Subjective and Objective) [199]. We believe that subjective methodologies encompass a) Replicating Swanson’s discoveries, and b) Incorporating expert opinion. The objective methodologies, on the other hand, encompass a) Using statistical evaluation methods, and b) Publishing in the medical domain. All the proposed evaluation methodologies, however, are subject to the following drawbacks:

<u>Generalizability</u>

- Replicating Swanson’s discoveries may introduce bias into systems’ design and does not guarantee systems’ generalizability into different cases.
- In incorporating experts’ opinions, different experts may not reach a consensus about the validity and interestingness of a specific discovery.
- Current evaluation metrics are inclined towards IR metrics and probabilistic approach [199]. But our observation based on the literature survey shown earlier revealed that at least 11 different LBD approaches exist of which the probabilistic approach is just one of them.
- Automated evaluation methodology [2] is conducted via the authors’ system such as LitLinker. It makes sense to test the performance of different correlation measures on the same system platform. However, no one can confidently assert and generalize the winning measure’s discovery performance because its evaluation is closely tight to the specific features of the platform (e.g. LitLinker). For instance, LitLinker represents each medical document using a set of its indexed MeSH terms. But Kostoff highlighted the problem associated with the fallibility of the human indexer (i.e. the Indexer Effect) and argued that any potential discovery made using a MeSH-based process must be validated not only in MeSH space but in text (i.e. the un-indexed words) space as well [198]. Further, we notice that the evaluation methodology proposed in [2] is hardly a novel methodology. Rather, it is a mere extension of their previous works on LitLinker. In [168, 167], the authors have evaluated the performance of two correlation measures, mutual information and z-score, using an evaluation methodology [167] that has very little difference from the evaluation methodology proposed in by [2]. The latter merely included two additional correlation measures, tf-idf and association rules, into the evaluation. The evaluation method is not as new as the author claim.

<u>Quality of gold standards</u>

- LBD attempts to model the structure of the scientific literature, not of nature. A crucial challenge with gold standards that has escaped the attention of LBD researchers: not all knowledge is discoverable from the literature. Some discoveries come purely from experimental data, direct observations of nature, or simply a pure chance for which there are no contributing evidence from the literature. In short, they are not inferable by LBD system. Therefore, the process of establishing a gold standard must demonstrate that it is sufficiently inferable from the literature by the LBD systems. Interestingly, to our knowledge, no attempt or measure has been made to address this challenge.
- Kostoff et al demonstrated that some of the existing gold standards, in the absence of a rigorous vetting procedure, are not genuine scientific discoveries [161].
- Yetisgen-Yildiz and Pratt [2] construct their gold standard from target terms that co-occur with the starting term in future literature set. Apart from applying additional semantic type filtering on these target terms, no further validation process is applied. Considering that two terms may co-occur for various reasons, these target terms cannot be a gold standard!
- Expert opinions are hard, if not impossible, to quantify. As a result, such a gold standard cannot be used to compare different systems.

The missing middle path between two extremes An LBD evaluation methodology cannot be formulated as a completely objective test because true scientific discoveries have an intricate set of criteria that should be satisfied such as novelty, relevance, non-triviality, validity, verifiability, simplicity, actionability, meaningfulness, etc. [1]. These criteria can only be determined by a consensus of human experts which, in effect, introduces subjectivity into the evaluation process. It cannot be left as an entirely subjective endeavour, either. To a certain extent, the evaluation method must be objective to ensure its generalizability. A middle path must be struck between the two extremes. For a start, we identify a similar paradigm underlying two dominant methods from both ends: (1) the replication of Swanson's discoveries and (2) the statistical methods. We call this paradigm the retrospective paradigm. Wren [198] has highlighted the feasibility of this approach. Retrospective paradigm uses historical data to predict known 'future' discoveries. If the average prediction accuracy of an LBD system is considerably good, it is reasonable to assume that it will also produce considerably reliable results in predicting the unknown future discoveries based on the current data. In both method (1) and (2), the paradigm is evident from the usage of specific cut-off dates for obtaining literature sets before and after the target discoveries. Since the retrospective paradigm is found to be operational in both extremes, it is plausible to conclude that it accommodates both subjective and objective evaluation elements, making it a suitable ground for carving the middle path.

Given the retrospective paradigm, how should the objective and subjective evaluation elements be combined? Two important components in machine learning system evaluations are: corpus and metrics. The corpus constitutes the gold standard of LBD system evaluation. It is possible for a group of domain experts who are independent from the creators of the LBD systems to curate the corpus. This ensures objectivity. Their selection of a set of valid scientific discoveries into the corpus as the gold standard ensures that subjective qualities of discoveries, as stated above, are fully or partially satisfied. It is interesting to see how such corpus is mostly non-existent in most LBD evaluations with the exception of a small set of 'gold standards' used by [190]. Quantitative evaluation metrics are inherently objective. The most difficult problem with this is to associate the metrics with a set of discovery qualities which are mostly subjective and qualitative.

The problem with ranking Both of the existing subjective (i.e. replication of Swanson's discoveries) and objective (i.e. statistical) evaluation methods have been entirely dominated by the IR-centric ranking evaluation paradigm. Factors leading to this are quite easy to understand: (a) the view of LBD as a subset (or superset!) of the IR problem. This view is supported by our literature review indeed, and (b) the inherent trade-off between systems' recall and precision for which the ranking scheme provides a convenient scoring mechanism.

We do not dismiss the usefulness of the ranking evaluation scheme. LBD systems are likely to generate many candidate target discoveries. Ranking these candidates has a strong practical reason: to ease the users' cognitive load during evaluation. However, three problems are observed here: (a) many papers replicated Swanson's discoveries but ranked them considerably low in the list. This is not practical because in real-world scenarios, users cannot be reasonable expected to view (or scroll to) results residing at such a low rank. (b) many papers cannot evaluate the validity of the other findings at higher ranks simply because there are too many of them. (c) Consequently, the true performance of the systems cannot be assessed, and the relationship between the scoring system and scientific discoveries cannot (at least, have not) be established.

BlackBox problem LBD approaches are mostly built on bag-of-words and term co-occurrences. The derivation of hypothesis, thus, becomes opaque to the user (a Black Box). There is no clear explanation of how two terms are related. While we can strongly argue that such explanation is important to assess the validity and the acceptance of the findings by domain experts, the most recent automated evaluation methodology [2] does not encourage the transparency of the systems.

Evaluation methodologies exist [1], but a gold standard that can be used as a basis of the comparison is still absent [198]. Until 2006, Bekhuis observed that LBD evaluation almost always entails replicating Swanson's earliest findings [200]. Bekhuis criticized the LBD community at the time as being too respectful of Swanson's methods. Weeber also noted that Swanson's original discovery was serendipitous [198]. In other words, it was not initially driven by a systematic process of scientific inquiry. Making Swanson's original discovery into an evaluation gold standard for LBD system is, at least in principle, not ideal. Kostoff highlighted,

". . . questions as to whether Swanson's hypotheses are true discoveries or are really innovations, and in any case his results give no indication of the extent of discoveries possible"(Bruza and Weeber, 2008).

Some evaluation approaches are highly subjective, focusing on a few predetermined (e.g. medical) discoveries as targets. Such evaluations may be biased towards desired results. In fact, Kostoff et al have demonstrated that discoveries claimed by these authors were not true scientific discoveries because the prior art actually existed. On the other extreme, some methodologies are entirely non-subjective [2, 30] that mere co-occurrences of terms in the future are regarded as discoveries without domain expert validation [161].

Most LBD systems utilize a scoring system and rank their results based on these scores. It is, therefore, natural to adopt the Recall-Precision evaluation paradigm [199]. Success is frequently judged as long as the target discovery is successfully recovered in the list. However, the target discovery is often found at the low ranking in the list and there is no attempt to evaluation findings at the top ranks. Hence, the actual effectiveness of the systems is not justified. Based on our research, a successful evaluation methodology has to address the following:

- Generalizable
  - i. Unbiased to a specific LBD approach
  - ii. Unbiased to an individual domain expert
- Quality gold standards
  - i. Corpus
    1. Inferable gold standards
    2. Setting the cut-off dates
    3. Format and representations
  - ii. Metrics. Meta-analysis: linking metrics to the qualities of scientific discoveries
- Integrate subjective and objective elements which, in turn, encompasses both Corpus and Metrics as well.
- Alternative evaluation methods. The ranking-based evaluation method remains practical and relevant because as far as we can see, LBD systems are likely to produce many candidate target discoveries and therefore need to rank them. It is plausible, however, to create alternative evaluation methods whose results may lead to more accurate LBD systems' ranking mechanisms.
  - i. Goodness of path. It is conceivable that two LBD systems link A and C through very different logical paths, documents, keywords, etc. A better LBD system should choose a 'better' path. What constitutes the 'better' path is an important research question.
  - ii. Early Discovery. Given the same target discovery, how early (over a time-line) can an LBD system discover it? The best LBD system should predict the target discovery earlier than its competitors.
  - iii. Noise discrimination. Given a target discovery buried in artificially-manipulated competing noises, can an LBD system reliably recover/detect it? The evaluation scenario may involve a series of test data, each of which is polluted with different amount/quality of noises. A better system should be able to recover the target discovery despite substantial amount of noises.

6. Conclusion and Future Work

Discovery in science is the result of the formulation of novel, interesting, and scientifically sensible hypotheses. These hypotheses can be formulated by reviewing the existing body of domain-specific literature. The voluminous amount of data stored in the literature, however, makes the task impossible to be performed manually by scientists. In this paper a modern classification of the existing LBD proposals is given. one of the observations is that amongst the different LBD approaches, only one of which is entirely objective relying on a probabilistic approach. Although it is quite dominant in the literature that LBD is seen as a sub-specialization of IR problem, we believe that there are important differences between the two problem domains with regards to novelty, time factor, reasoning, and relevance.

The paper also discusses an interesting topic that investigates the early indicators of relatedness. It is possible to think that the sudden emergence of publications concerning some concepts at roughly the same time could have been a natural response to a specific, significant event (e.g. a prior scientific discovery, a discovery of a new drug, etc.). This common response may serve as an early sign of their relatedness (which became obvious only decades later). However, this requires more research to be proved and regarded as a major stream for our future research. As Kostoff stated, "A central problem with all the LBD studies that have been reported in the open literature is the absence of a gold standard that can be used as a basis of comparison" [198]. Kostoff also highlighted, "Currently, it is not at all clear which LBD approaches are most efficient due to a lack of quantitative methods and gold standard test sets for analysis" [198]. Evaluating an LBD proposal is challenging and many proposals failed to prove objectivity. Thus, this paper proposes gold standards that could be used to evaluate LBD proposals avoiding subjectivity. The future of this research is centred on the proposal of a complete methodology that evaluate current LBD proposals and stand still for the future proposals as well.

References

[1] N.R. Smalheiser, JASIST 63(2), 218 (2012)
[2] M. Yetisgen-Yildiz, W. Pratt, Journal of Biomedical Informatics 42(4), 633 (2009)
[3] D.R. Swanson, N.R. Smalheiser, Artif. Intell. 91(2), 183 (1997)
[4] A. Singhal, M. Kaszkiel, in WWW (2001), pp. 708–716
[5] M.D. Gordon, R.K. Lindsay, JASIS 47(2), 116 (1996)
[6] M. Gordon, R. Lindsay, W. Fan, ACM Trans. Internet Techn. 2(4), 261 (2002)
[7] R.K. Lindsay, M.D. Gordon, JASIS pp. 574–587 (1999)
[8] H.W. Chun, Y. Tsuruoka, J.D. Kim, R. Shiba, N. Nagata, T. Hishiki, J. Tsujii, in Proceedings of the Pacific Symposium on Biocomputing (PSB) 11 (Maui, Hawaii, USA, 2006), pp. 4–15
[9] Y. Tsuruoka, J. Tsujii, S. Ananiadou, Bioinformatics 24(21), 2259 (2008)

[10] Y. Tsuruoka, M. Miwa, K. Hamamoto, J. Tsujii, S. Ananiadou, Bioinformatics 27(13), i111 (2011). DOI 10.1093/bioinformatics/btr214
[11] L.A. Adamic, D.M. Wilkinson, B.A. Huberman, E. Adar, in CSB (2002), pp. 109–117
[12] K. Sriphaew, A study on document relation discovery using frequent itemset mining. Ph.D. thesis, Information Technology Program Sirindhorn International Institute of Technology Thammasat University (2007)
[13] K. Sriphaew, T. Theeramunkong, IEICE Transactions on Information and Systems 90-D(8), 1225 (2007)
[14] T. Theeramunkong, K. Sriphaew, in NSTDA Annual Conference Science (NSTDA, 2007)
[15] K. Sriphaew, T. Theeramunkong, in CIDM (IEEE, 2007), pp. 293–299
[16] R. Frijters, S. Verhoeven, W. Alkema, R. van Schaik, J. Polman, Pharmacogenomics 8, 1521 (2007)
[17] R. Frijters, M. van Vugt, R. Smeets, R.C. van Schaik, J. de Vlieg, W. Alkema, PLoS Computational Biology 6(9) (2010)
[18] B.T.F. Alako, A. Veldhoven, S. van Baal, R. Jelier, S. Verhoeven, T. Rullmann, J. Polman, G. Jenster, BMC Bioinformatics 6, 51 (2005)
[19] B. Stapley, G. Benoit, Pac Symp Biocomput (2000)
[20] J.D. Wren, The iridescent system: an automated data mining method to identify, evaluate, and analyze sets of relationships within textual databases. Ph.D. thesis, The University of Texas Southwestern Medical Center at Dallas (2003). AAI0805307
[21] J. Wren, BMC Bioinformatics 5, 145 (2004)
[22] J.D. Wren, R. Bekeredjian, J.A. Stewart, R.V. Shohet, H.R. Garner, Bioinformatics 20(3), 389 (2004)
[23] T. Cohen, AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium pp. 126–130 (2008)
[24] T. Cohen, R.W. Schvaneveldt, Studies in health technology and informatics 160(Pt1), 661 (2010)
[25] T. Cohen, R. Schvaneveldt, D. Widdows, J. of Biomedical Informatics 43(2), 240 (2010). DOI 10.1016/j.jbi.2009.09.003
[26] T. Cohen, Widdows, D., S.R. W., R.T. C, in AAAI-Fall 2010 Symposium on Quantum Informatics for Cognitive, Social, and Semantic Processes (2010), p. 1113
[27] T. Cohen, G.K. Whitfield, R.W. Schvaneveldt, K. Mukund, T. Rindflesch, Journal of biomedical discovery and collaboration 5, 21 (2010)
[28] T. Cohen, R. Schvaneveldt, D. Widdows, J. of Biomedical Informatics 43(2), 240 (2010). DOI 10.1016/j.jbi.2009.09.003
[29] T. Cohen, D. Widdows, R.W. Schvaneveldt, P. Davies, T.C. Rindflesch, Journal of Biomedical Informatics 45(6), 1049 (2012)
[30] R. Schvaneveldt, T. Cohen, in Computer-Based Diagnostics and Systematic Analysis of Knowledge, ed. by D. Ifenthaler, P. Pirnay-Dummer, N.M. Seel (Springer US, 2010), pp. 189–211. DOI 10.1007/978-1-4419-5662-0\ 11
[31] D. Widdows, P. Bruza, in AAAI Spring Symposium: Quantum Interaction (AAAI, 2007), pp. 126–133
[32] D. Widdows, T. Cohen, in Proceedings of the 2010 IEEE Fourth International Conference on Semantic Computing (IEEE Computer Society, Washington, DC, USA, 2010), ICSC '10, pp. 9–15. DOI 10.1109/ICSC.2010.94
[33] M.D. Gordon, S. Dumais, J. Am. Soc. Inf. Sci. 49(8), 674 (1998). DOI 10.1002/(SICI) 1097-4571(199806)49:8h674::AID-ASI2i3.0.CO;2-Q
[34] F.A.D. Neves, E.A. Fox, X. Yu, in CIKM (2005), pp. 91–98
[35] E.A. Fox, F.A.D. Neves, X. Yu, R. Shen, S. Kim, W. Fan, Commun. ACM 49(4), 52 (2006)
[36] A. Sehgal, A. Sehgal, X.Y. Qiu, P. Srinivasan, in In Proceedings of the SIGIR 2003 Workshop on Text Analysis and Search for Bioinformatics (2003)
[37] A.K. Sehgal, Exploring concept spaces for text mining. Ph.D. thesis, Department of Computer Science in the Graduate College of The University of Iowa (2003)
[38] A.K. Sehgal, P. Srinivasan, in SIGIR, ed. by R.A. Baeza-Yates, N. Ziviani, G. Marchionini, Moffat, J. Tait (2005)
[39] A.K. Sehgal, P. Srinivasan., in I3 (2007)
[40] P. Srinivasan, J. Am. Soc. Inf. Sci. Technol. 55(5), 396 (2004). DOI 10.1002/asi.10389
[41] P. Srinivasan, B. Libbus, Bioinformatics 20(Supp 1), i290 (2004)
[42] P. Srinivasan, B. Libbus, A.K. Sehgal, in Proceedings of the Workshop BioLINK, Linking Biological Literature, Ontologies and Databases at HLT NAACL, ed. By L. Hirschman, J. Pustejovsky (Association for Computational Linguistics, Boston, Massachusetts, USA, 2004), pp. 33–40
[43] S. Lee, K. Park, D. Kim, Expert Opinion on Drug Discovery 4(11), 1 (2009)
[44] S. Lee, J. Choi, K. Park, M. Song, D. Lee, in Proceedings of the ACM fifth international workshop on Data and text mining in biomedical informatics (ACM, New York, NY, USA, 2011), DTMBIO '11, pp. 27–34. DOI 10.1145/2064696.2064704
[45] S. Lee, J. Choi, K. Park, M. Song, L. Doheon, BMC Medical Informatics and Decision Making 12(1), 1 (2012). DOI 10.1186/1472-6947-12-S1-S1
[46] E.M. Van Mulligen, C. Van Der Eijk, J.A. Kors, B.J. Schijvenaars, B. Mons, Proceedings / AMIA ... Annual Symposium. AMIA Symposium pp. 835–839 (2002)
[47] C.C. van der Eijk, E.M. van Mulligen, J.A. Kors, B. Mons, J. van den Berg, J. Am. Soc. Inf. Sci. Technol. 55(5), 436 (2004). DOI 10.1002/asi.10392
[48] R. Jelier, G. Jenster, L.C.J. Dorssers, C.C. Van Der Eijk, E.M. Van Mulligen, B. Mons, J.A. Kors, Bioinformatics 21(9), 2049 (2005). DOI 10.1093/bioinformatics/bti268
[49] R. Jelier, Text mining applied to molecular biology. Ph.D. thesis, Erasmus MC: University Medical Center Rotterdam (2008)
[50] R. Jelier, G. Jenster, L. Dorssers, B. Wouters, P. Hendriksen, B. Mons, R. Delwel, J. Kors, BMC Bioinformatics 8(1), 14 (2007). DOI 10.1186/1471-2105-8-14
[51] R. Jelier, M.J. Schuemie, A. Veldhoven, L.C. Dorssers, G. Jenster, J.A. Kors, Genome biology 9(6) (2008). DOI 10.1186/gb-2008-9-6-r96
[52] M.J.S. Jelier, R., P.J. Roes, E.M. van Mulligen, J.A. Kors, International Journal of Medical Informatics 77(5), 354 (2008)

[53] R. Jelier, J.J. Goeman, K.M. Hettne, M.J. Schuemie, J.T. den Dunnen, P.A.C. ’t Hoen, Briefings in Bioinformatics 12(5), 518 (2011)
[54] W.M.L.M.t.C.H.B.S.K.J.L.B. Hettne, KM., JOURNAL OF CLINICAL PERIODONTOLOGY 34 (12), 1016 (2007)
[55] K.M. Hettne, E.M. van Mulligen, M.J. Schuemie, B.J.A. Schijvenaars, J.A. Kors, J. Biomedical Semantics 1, 5 (2010)
[56] H. van Haagen, P. ’t Hoen, A. Botelho Bovo, A. de Morr´ee, E. van Mulligen, C. Chichester, J. Kors, J. den Dunnen, G. van Ommen, S. van der Maarel, V. Kern, B. Mons, M. Schuemie, PloS one 4(11) (2009)
[57] van Haagen, ’t Hoen PA, de Morre A, van Roon-Mom WM, P. DJ, R. M, M. B, van Ommen GJ, S. MJ., Proteomics. 11(5), 843 (2011)
[58] M. Chagoyen, P. Carmona-Saez, H. Shatkay, J.M. Carazo, A.D. Pascual-Montano, BMC Bioinformatics 7, 41 (2006)
[59] H.S. Blaschke, C.; L. Hirschman, A. Valencia, in Proc. of the Workshop on Linking Literature, Information and Knowledge in Bi- ology (BioLINK)., ed. by L.N. in Bioinformatics (2010), 6004
[60] van Driel MA, J. Bruggeman, G. Vriend, H. Brunner, J. Leunissen, Eur J Hum Genet. 15 (5), 535 (2006)
[61] T. Urbancic, I. Petric, B. Cestnik, M. Macedoni-Luksic, in AIME (2007), pp. 217–226
[62] T. Urbancic, I. Petric, B. Cestnik, in ISMIS (2009), pp. 129–138
[63] I. Petric, T. Urbancic, B. Cestnik, M. Macedoni-Luksic, Journal of Biomedical Informatics 42(2), 219 (2009)
[64] I. Petric, B. Cestnik, N. Lavrac, T. Urbancic, Comput. J. 55(1), 47 (2012)
[65] M. Jurˇsiˇc, N.L. Igor Mozetiˇc, Miha Grˇcar, Bisociative knowledge discovery via bterm identification. Tech. rep., Department of Knowledge Technologies Joˇz ef Stefan Institute Jamova 39, 1000 Ljubljana Slovenija (2010)
[66] M. Jurˇsiˇc, B. Cestnik, T. Urbanˇciˇc, N. Lavraˇc, in Proceedings of the 3rd International Conference on Computational Creativity, ed. by M.L. Maher, K. Hammond, A. Pease, R. P´erez y P´erez, D. Ventura, G. Wiggins (University College Dublin, Dublin, Ireland, 2012), pp. 33–40
[67] M. Jurˇsiˇc, B. Sluban, B. Cestnik, M. Grˇcar, N. Lavraˇc, in Bisociative Knowledge Discovery, Lecture Notes in Computer Science, vol. 7250, ed. by M. Berthold (Springer Berlin Heidelberg, 2012), pp. 66–90. DOI 10.1007/978-3-642-31830-6\ 6
[68] M. Jurˇsiˇc, B. Cestnik, T. Urbanˇciˇc, N. Lavraˇc, in Bisociative Knowledge Discovery, Lecture Notes in Computer Science, vol. 7250, ed. by M. Berthold (Springer Berlin Heidelberg, 2012), pp. 338–358. DOI 10.1007/978-3-642-31830-6\ 24
[69] A. Koike, T. Takagi, Journal of the American Society for Information Science and Technology 58(1), 51 (2007). DOI 10.1002/asi.20421
[70] A. Koike, in Literature-based Discovery, Information Science and Knowledge Management, vol. 15, ed. by P. Bruza, M. Weeber (Springer Berlin Heidelberg, 2008), pp. 173–192. DOI 10.1007/978-3-540-68690-3\ 11
[71] H. Shuiqing, L.H.B. Yang, M. Zhang (eds.). A compound correlation model for disjoint literature-based knowledge discovery, no. 423 - 436 in 4 (Emerald Group Publishing Limited, 2012). DOI 10.1108/00012531211244770
[72] W. Maciel, A. Faria-Campos, M. Gonalves, S. Campos, BMC Genomics 12(4), 1 (2011). DOI 10.1186/1471-2164-12-S4-S1
[73] D. Song, P.D. Bruza, J. Am. Soc. Inf. Sci. Technol. 54(4), 321 (2003). DOI 10.1002/asi.10213
[74] D. Song, P. Bruza, in Frontiers of WWW Research and Development – APWeb 2006, Lecture Notes in Computer Science, vol. 3841, ed. by X. Zhou, J. Li, H. Shen, M. Kitsuregawa, Y. Zhang (Springer Berlin Heidelberg, 2006), pp. 692–701. DOI 10.1007/11610113\ 60
[75] D. Song, M. Lalmas, K. van Rijsbergen, I. Frommholz, B. Piwowarski, J. Wang, P. Zhang, G. Zuccon, P.D. Bruza, S. Arafat, L. Azzopardi, E.D. Buccio, A. Huertas- Rosero, Y. Hou, M. Melucci, S. Ruger, in AAAI Fall Symposium on Quantum Informatics for Cognitive, Social and Semantic Processes 2010, ed. by P.D. Bruza, W. Lawless, K. van Rijsbergen, D.A. Sofge (AAAI Press, Arlington, Va, 2010), pp. 105–108
[76] D. Song, P. Bruza, R. McArthur, Logic Journal of IGPL Vol. 12(No. 2), 97 (2004)
[77] R. Cole, P. Bruza, in Discovery Science, Lecture Notes in Computer Science, vol. 3735, ed. by A. Hoffmann, H. Motoda, T. Scheffer (Springer Berlin Heidelberg, 2005), pp. 84–98. DOI 10.1007/11563983\ 9
[78] P. Bruza, R. Cole, D. Song, Z. Bari, Logic Journal of the IGPL 14(2), 161 (2006)
[79] P.D. Bruza, in Information Symposium on Information Technology (ITSim ‘08) (IEEE, Kuala Lumpur, 2008), pp. 1–9. (keynote presentation paper)
[80] P. Bruza, K. Kitto, B. Ramm, L. Sitbon, D. Song, in Proceedings of Model-based Reasoning Conference in Science and Technology (Campinas, Brazil, 2009)
[81] P.D. Bruza, K. Kitto, B.J. Ramm, L. Sitbon, D. Song, S. Blomberg, Logic Journal of the IGPL 20(2), 445 (2012)
[82] P.D. Bruza, K. Kitto, B. Ramm, L. Sitbon, D. Song, S. Blomberg, Logic Journal of IGPL 20(2), 445 (2012). DOI 10.1093/jigpal/jzq049
[83] D. Aerts, P. Bruza, Y. Hou, J. Jose, M. Melucci, J.Y. Nie, D. Song, Procedia Computer Science 7, 278 (2011)
[84] S. Aerts, K. Kitto, L. Sitbon, in Quantum Interaction, Lecture Notes in Computer Science, vol. 7052, ed. by D. Song, M. Melucci, I. Frommholz, P. Zhang, L.Wang, S. Arafat (Springer Berlin Heidelberg, 2011), pp. 13–24. DOI 10.1007/978-3-642-24971-6\ 3
[85] S. Dar´anyi, P. Wittek, in QI (2012), pp. 207–217
[86] R. Homayouni, K. Heinrich, L. Wei, M.W. Berry, Bioinformatics 21(1), 104 (2005). DOI 10.1093/bioinformatics/bth464
[87] S. Roy, K. Heinrich, V. Phan, M.W. Berry, R. Homayouni, BMC Bioinformatics 12(S-10), S19 (2011)
[88] E. Tjioe, M. Berry, R. Homayouni, BMC Bioinformatics 11(Suppl 6), 1 (2010). DOI 10.1186/1471-2105-11-S6-S14
[89] J. Stegmann, G. Grohmann, in Literature-based Discovery, Information Science and Knowledge Management, vol. 15,

ed. by P. Bruza, M. Weeber (Springer Berlin Heidelberg, 2008), pp. 115–131. DOI 10.1007/978-3-540-68690-3\ 8
[90] N. Ford, Journal of the American Society for Information Science 42(1), 72 (1991). DOI 10.1002/(SICI)1097-4571(199101)42:1h72::AID-ASI8i3.0.CO;2-9
[91] M. Weeber, H. Klein, A.R. Aronson, J.G. Mork, L.T. de Jong-van den Berg, R. Vos, Proceedings / AMIA ... Annual Symposium. AMIA Symposium pp. 903–907 (2000)
[92] M. Weeber, H. Klein, L.T.W. de Jong-van den Berg, R. Vos, JASIST 52(7), 548 (2001)
[93] M.Weeber, R. Vos, H. Klein, L.T.W. de Jong-van den Berg, A.R. Aronson, G. Molema, JAMIA 10(3), 252 (2003)
[94] M. Weeber, in Computational Discovery of Scientific Knowledge (2007), pp. 290–306
[95] L.J. Jensen, P. Bork, PLoS Biol 8(5), e1000374+ (2010). DOI 10.1371/journal.pbio. 1000374
[96] T. Miyanishi, K. Seki, K. Uehara, in SAC, ed. by S.Y. Shin, S. Ossowski, M. Schumacher, M.J. Palakal, C.C. Hung (ACM, 2010), pp. 1552–1558
[97] T. Miyanishi, S. Kazuhiro, U. Kuniaki, IPSJ Transactions on Bioinformatics(TBIO) 4, 9 (2011)
[98] C. Perez-Iratxeta, M. Wjst, P. Bork, M.A. Andrade, BMC Genet 6(1) (2005). DOI 10.1186/1471-2156-6-45
[99] J.O. Korbel, T. Doerks, L.J. Jensen, C. Perez-Iratxeta, S. Kaczanowski, S.D. Hooper, M.A. Andrade, P. Bork, PLoS Biol 3(5), e134+ (2005). DOI 10.1371/journal.pbio.0030134
[100] C. Perez-Iratxeta, P. Bork, M.A. Andrade-Navarro, Nucl. Acids Res. 35(suppl 2), W212 (2007). DOI 10.1093/nar/gkm223
[101] N.C. Baker, B.M. Hemminger, J. of Biomedical Informatics 43(4), 510 (2010). DOI 10.1016/j.jbi.2010.03.008
[102] N. Tiffin, J.F. Kelso, A.R. Powell, H. Pan, V.B. Bajic, W.A. Hide, Nucleic acids research 33(5), 1544 (2005). DOI 10.1093/nar/gki296
[103] J.D. Kim, T. Ohta, J. ichi Tsujii, BMC Bioinformatics 9 (2008)
[104] J. Hur, A.D. Schuyler, D.J. States, E.L. Feldman, Bioinformatics. 25(16), 838840 (2009)
[105] J. Hur, K.A. Sullivan, A.D. Schuyler, Y. Hong, M. Pande, D.J. States, H.V. Jagadish, E.L. Feldman, BMC medical genomics 3(1) (2010). DOI 10.1186/1755-8794-3-49
[106] J. Hur, A. ¨Ozg¨ur, Z. Xiang, Y. He, J. Biomedical Semantics 3, 18 (2012)
[107] A. ¨Ozg¨ur, D.R. Radev, in Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3 (Association for Computational Linguistics, Stroudsburg, PA, USA, 2009), EMNLP ’09, pp. 1398–1407
[108] A. ¨Ozg¨ur, H.Y. Xiang Z, Radev DR, Journal of Biomedicine and Biotechnology 2010 (2010). DOI 10.1155/2010/426479
[109] J.R. Katukuri, Y. Xie, V.V. Raghavan, in BIBM (2009), pp. 366–370
[110] J.R. Katukuri, Y. Xie, V.V. Raghavan, A. Gupta, in BIBM (2011), pp. 562–568
[111] J. Katukuri, Y. Xie, V. Raghavan, A. Gupta, BMC Genomics 13(Suppl 3), S5 (2012). DOI 10.1186/1471-2164-13-S3-S5
[112] J.P. Arrais, J.G. Rodrigues, J.L. Oliveira, in 2nd International Workshop on Practical Applications of Computational Biology and Bioinformatics (Iwpacbb 2008) (Springer Berlin Heidelberg, 2009), pp. 74–82
[113] J. Arrais, J. Oliveira, in Information Technology and Applications in Biomedicine (ITAB), 2010 10th IEEE International Conference on (2010), pp. 1–4. DOI 10.1109/ITAB.2010.5687629
[114] J.P. Arrais, J.L. Oliveira, Open Access Bioinformatics 3, 123 (2011)
[115] Y. Li, P. Agarwal, PLoS ONE 4(2), e4346+ (2009). DOI 10.1371/journal.pone.0004346
[116] L. Eronen, H. Toivonen, BMC Bioinformatics 13, 119 (2012)
[117] B. Wilkowski, Semantic approaches for knowledge discovery and retrieval in biomedicine. Ph.D. thesis, Department of Informatics and Mathematical Modelling, Technical University of Denmark (2011)
[118] B. Wilkowski, M. Fiszman, C.M. Miller, D. Hristovski, S. Arabandi, G. Rosemblat, T.C. Rindflesch, AMIA Annual Symposium Proceedings 2011, 1514 (2011)
[119] G. Gonzalez, L. Tari, A. Gitter, R. Leaman, S. Nikkila, R. Wendt, A. Zeigler, C. Baral, Proceedings of the Second BioCreative Challenge Workshop-Critical Assessment of Information Extraction in Molecular Biology (2007)
[120] G. Gonzalez, J. URIBE, L. TARI, C. BROPHY, C. BARAL, in Pacific Symposium on Biocomputing, vol. 12 (2007), vol. 12, pp. 28–39
[121] L. Tari, S. Anwar, S. Liang, J. Cai, C. Baral, Bioinformatics 26(18), i547 (2010)
[122] L. Tari, S. Anwar, S. Liang, J. Hakenberg, C. Baral, Proc. of PSB pp. 465–476 (2010)
[123] L. Tari, N. Vo, S. Liang, J. Patel, C. Baral, J. Cai, PloS one 7(7), e40946+ (2012). DOI 10.1371/journal.pone.0040946
[124] J. Hakenberg, Mining relations from the biomedical literature. Ph.D. thesis, Humboldt University of Berlin (2009)
[125] J. Li, X. Zhu, J.Y. Chen, in Proceedings of the 2008 ACM symposium on Applied computing (ACM, New York, NY, USA, 2008), SAC ’08, pp. 1287–1291. DOI 10.1145/1363686.1363984
[126] J. Li, Z. Xiaoyan, C.J. Yue, PLoS Computational Biology 5(7) (2009)
[127] L. Jiao, Z. Xiaoyan, C.J. Yue, IJDMB 4(3), 241 (2010)
[128] T. Li, B. Song, Z. Wu, M. Lu, W.G. Zhu, Briefings in Bioinformatics (2013). DOI 10.1093/bib/bbt060
[129] X. Hu, X. Zhang, I. Yoo, X. Wang, J. Feng, Int. J. Intell. Syst. 25(2), 207 (2010). DOI 10.1002/int.v25:2
[130] H.R. Turtle, W.B. Croft, Comput. J. 35(3), 279 (1992)
[131] K. Seki, J. Mostafa, Pacific Symposium on Biocomputing 12, 316 (2007)
[132] S. Kazuhiro, J. Mostafa, in Discovery Science, Lecture Notes in Computer Science, vol. 4755, ed. by V. Corruble,

M. Takeda, E. Suzuki (Springer Berlin Heidelberg, 2007), pp. 185–196. DOI 10.1007/978-3-540-75488-6\ 18
[133] C. Chen, Journal of the American Society for Information Science and Technology 63(3), 431 (2012). DOI 10.1002/asi.21694
[134] C. Chen, Y. Chen, M. Horowitz, H. Hou, Z. Liu, D. Pellegrino, J. Informetrics 3(3), 191 (2009)
[135] J. Zhang, M.S.E. Vogeley, C. Chen, Scientometrics 86(1), 1 (2011)
[136] S. Zadrozny, K. Nowacka, Fuzzy Sets and Systems 160(15), 2173 (2009)
[137] J.D. Wren, Soft Comput. 10(4), 374 (2006)
[138] C. Perez-Iratxeta, P. Bork, M.A. Andrade, Nature genetics 31(3), 316 (2002). DOI 10.1038/ng895
[139] D. Shahaf, C. Guestrin, E. Horvitz, in Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (ACM, New York, NY, USA, 2012), KDD ’12, pp. 1122–1130. DOI 10.1145/2339530.2339706
[140] D. Kumar, R.F.H. Malcolm Potts, Naren Ramakrishnan, in Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD ’06) (2006), pp. 604–610
[141] M.S. Hossain, Exploratory data analysis using clusters and stories. Ph.D. thesis, The Faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of, Blacksburg, Virginia (2012)
[142] M.S. Hossain, J. Gresock, Y. Edmonds, R. Helm, M. Potts, N. Ramakrishnan, PloS one 7(1), e29509+ (2012). DOI 10.1371/journal.pone.0029509
[143] J. Gresock, D. Kumar, R. Helm, M. Potts, N. Ramakrishnan, Mining novellas from pubmed abstracts using a storytelling algorithm. Departmental Technical Report TR- 07-08, Department of Computer Science, Virginia Tech. (2007)
[144] D. Kumar, N. Ramakrishnan, R. Helm, M. Potts, Knowledge and Data Engineering, IEEE Transactions on 20(6), 736 (2008). DOI 10.1109/TKDE.2008.32
[145] M.S. Hossain, M. Narayan, N. Ramakrishnan, CoRR abs/1002.3195 (2010)
[146] X. Liu, W.B. Croft, in Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval (ACM, New York, NY, USA, 2004), SIGIR ’04, pp. 186–193. DOI 10.1145/1008992.1009026
[147] R.N. Kostoff, H.J. Eberhart, D.R. Toothman, Journal of Information Science 23(4), 301 (1997). DOI 10.1177/016555159702300404
[148] R.N. Kostoff, T. Braun, A. Schubert, D.R. Toothman, J.A. Humenik, Journal of Chemical Information and Computer Sciences 40(1), 19 (2000)
[149] P.B. Losiewicz, D.W. Oard, R.N. Kostoff, J. Intell. Inf. Syst. 15(2), 99 (2000)
[150] R.N. Kostoff, in ISMIS (2000), pp. 86–96
[151] R.N. Kostoff, J.A. del R´ıo, J.A. Humenik, E.O. Garc´ıa, A.M. Ram´ırez, JASIST 52(13), 1148 (2001)
[152] R.N. Kostoff, in Discovery Science (2001), pp. 196–213
[153] R.N. Kostoff, M.F. Shlesinger, R. Tshiteya, I. J. Bifurcation and Chaos 14(1), 61 (2004)
[154] R.N. Kostoff, J.A. Block, J.A. Stump, K.M. Pfeil, I. J. Medical Informatics 73(6), 515 (2004)
[155] R.N. Kostoff, J.A. Block, JASIST 56(9), 946 (2005)
[156] R.N. Kostoff, J.T. Rigsby, R.B. Barth, J. Information Science 32(6), 581 (2006)
[157] R.N. Kostoff, Journal of Biomedical Informatics 40(4), 448 (2007)
[158] R. Kostoff, Scientometrics 72(3), 513 (2007)
[159] R.N. Kostoff, J. Informetrics 2(4), 354 (2008)
[160] R.N. Kostoff, R.B. Barth, C.G.Y. Lau, Scientometrics 76(1), 43 (2008)
[161] R.N. Kostoff, J.A. Block, J.L. Solka, M.B. Briggs, R.L. Rushenberg, J.A. Stump, D. Johnson, T.J. Lyons, J.R. Wyatt, ARIST 43(1), 1 (2009)
[162] H. Chen, R.N. Kostoff, C. Chen, J. Zhang, M.S.E. Vogeley, K. B¨orner, N. Ma, R.J. Duhon, A. Zoss, V. Srinivasan, E.A. Fox, C.C. Yang, C.P. Wei, IEEE Intelligent Systems 24(4), 68 (2009)
[163] R.N. Kostoff, J. Information Science 36(1), 104 (2010)
[164] D.J. Schoeneck, A.L. Porter, R.N. Kostoff, E.M. Berger, Techn. Analysis & Strat. Manag. 23(6), 601 (2011)
[165] R.N. Kostoff, JASIST 63(8), 1675 (2012)
[166] W. Pratt, M. Yetisgen-Yildiz, in Proceedings of the 2nd international conference on Knowledge capture (ACM, New York, NY, USA, 2003), K-CAP ’03, pp. 105–112. DOI 10.1145/945645.945662
[167] M. Yetisgen-Yildiz, in Proceedings of 29th Annual International ACM SIGIR Conference on Research & Development on Information Retrieval (SIGIR’06) Doctoral Consortium (Seattle, WA, 2006)
[168] M. Yetisgen-Yildiz, W. Pratt, Journal of Biomedical Informatics pp. 600–611 (2006)
[169] D. Hristovski, J. Stare, B. Peterlin, S. Dzeroski, Medinfo 10(Pt 2), 1344 (2001)
[170] D. Hristovski, B. Peterlin, J.A. Mitchell, S.M. Humphrey, Stud Health Technol Inform 95, 68 (2003)
[171] D. Hristovski, B. Peterlin, J. Mitchell, S. Humphrey, I. J. Medical Informatics 74(2-4), 289 (2005)
[172] D. Hristovski, C. Friedman, T.C. Rindflesch, B. Peterlin, AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium pp. 349–353 (2006)
[173] D. Hristovski, C. Friedman, T. Rindflesch, B. Peterlin, in Literature-based Discovery, Information Science and Knowledge Management, vol. 15, ed. by P. Bruza, M. Weeber (Springer Berlin Heidelberg, 2008), pp. 133–152. DOI 10.1007/978-3-540-68690-3\ 9
[174] D. Hristovski, A. Kastrin, B. Peterlin, T.C. Rindflesch, AMIA Annu Symp Proc. p. 255259 (2009)
[175] D. Hristovski, A. Kastrin, B. Peterlin, T. Rindflesch, in Proceedings of the 2009 workshop of the BioLink Special Interest Group, international conference on Linking Literature, Information, and Knowledge for Biology (Springer-Verlag, Berlin, Heidelberg, 2010), ISMB/ECCB’09, pp. 53–61. DOI 10.1007/978-3-642-13131-8\ 7

[176] D. Hristovski, T. Rindflesch, B. Peterlin, Cardiovascular & hematological agents in medicinal chemistry 11(1), 14 (2013)
[177] P. Gandra, M. Pradhan, M.J. Palakal, in Proceedings of the International Symposium on Biocomputing (ACM, New York, NY, USA, 2010), ISB ’10, pp. 9:1–9:8. DOI 10.1145/1722024.1722035
[178] X. Hu, I. Yoo, M. Song, Y. Zhang, I.Y. Song, in Proceedings of the 14th ACM international conference on Information and knowledge management (ACM, New York, NY, USA, 2005), CIKM ’05, pp. 249–250. DOI 10.1145/1099554.1099611
[179] X. Hu, X. Xu, Int. J. Web Grid Serv. 1(2), 222 (2005). DOI 10.1504/IJWGS.2005. 008321
[180] X. Hu, G. Li, I. Yoo, X. Zhang, X. Xu, in GrC, ed. by X. Hu, Q. Liu, A. Skowron, T.Y. Lin, R. Yager, B. Zhang (IEEE, 2005), pp. 22–27
[181] X. Hu, X. Zhang, D. Wu, X. Zhou, P. Rumm, in Intelligence and Security Informatics, Lecture Notes in Computer Science, vol. 3975, ed. by S. Mehrotra, D. Zeng, H. Chen, B. Thuraisingham, F.Y. Wang (Springer Berlin Heidelberg, 2006), pp. 548–553. DOI 10.1007/11760146\ 55
[182] D.R. Swanson, N.R. Smalheisert, Link analysis of medlinteitles as an aid to scientific discovery. Aaai technical report fs-98-01, Division of Humanities Unniversity of Chicago and Department of Psychiatry, University of Illinois (1998)
[183] R. Swanson, Library Trends 48(1) (1999)
[184] D.R. Swanson, N.R. Smalheiser, A. Bookstein, JASIST 52(10), 797 (2001)
[185] D.R. Swanson, N.R. Smalheiser, V.I. Torvik, JASIST 57(11), 1427 (2006)
[186] N.R. Smalheiser, Technovation 21(10), 689 (2001). DOI 10.1016/s0166-4972(01)00048-7
[187] N. Smalheiser, in Discovery Science (2005), pp. 26–43
[188] N.R. Smalheiser, V.I. Torvik, W. Zhou, Computer Methods and Programs in Biomedicine 94(2), 190 (2009)
[189] N.R. Smalheiser, W. Zhou, V.I. Torvik, Information 2(2), 266 (2011)
[190] V.I. Torvik, N.R. Smalheiser, Bioinformatics 23(13), 1658 (2007)
[191] W. Huang, Y. Nakamori, in Fuzzy Information, 2004. Processing NAFIPS ’04. IEEE Annual Meeting of the, vol. 1 (2004), vol. 1, pp. 450–455 Vol.1. DOI 10.1109/NAFIPS. 2004.1336325
[192] W. Huang, Y. Nakamori, S. Wang, T. Ma, in Computational and Information Science, Lecture Notes in Computer Science, vol. 3314, ed. by J. Zhang, J.H. He, Y. Fu (Springer Berlin Heidelberg, 2005), pp. 794–799. DOI 10.1007/978-3-540-30497-5\ 123
[193] H. Wei, N. Yoshiteru, W. Shouyang, M. Tieju, Intell. Data Anal. 9(2), 219 (2005)
[194] W. Huang, S. Wang, L. Yu, H. Ren, in Knowledge Discovery in Life Science Literature, Lecture Notes in Computer Science, vol. 3886, ed. by E. Bremer, J. Hakenberg, E.H. Han, D. Berrar, W. Dubitzky (Springer Berlin Heidelberg, 2006), pp. 68–77. DOI 10.1007/11683568\ 6
[195] J. Stegmann, G. Grohmann, Scientometrics 56(1), 111 (2003)
[196] S. J., G. G., CoRR abs/cs/0509020 (2005)
[197] R.N. Kostoff, R.G. Koytcheff, C.G. Lau, in Data Mining Applications for Empowering Knowledge Societies (IGI Global, 2009), pp. 198–219. DOI 10.4018/978-1-59904-657-0.ch011
[198] P. Bruza, M. Weeber, Literature-based Discovery, vol. 15 (Springer Berlin Heidelberg, 2008)
[199] M. Yetisgen-Yildiz, wanda Pratt, in Literature-based Discovery, Information Science and Knowledge Management, vol. 15, ed. by P. Bruza, M. Weeber (Springer Berlin Heidelberg, 2008), pp. 101–113. DOI 10.1007/978-3-540-68690-3\ 7
[200] T. Bekhuis, Biomedical Digital Libraries 3(1), 2+ (2006). DOI 10.1186/1742-5581-3-2

Ali ahmed received the B.S. and M.S. degrees in Computer Science from Faculty of Computers and information in 2000 and 2004, respectively. During 2006-2010, he stayed in the Information and Management Group (IMG) at the University of Manchester, UK studying Computer security. He now an assistant professor at Cairo University and an Honorary Lecturer at the University of Liverpool, UK.