Narrowing gaps in science

Ulises Cerviño Beresi
The Robert Gordon University
School of Computing
Aberdeen, UK
ucb@comp.rgu.ac.uk

Abstract: As a consequence of the natural fragmentation of science into specialities, disjoint but logically related literatures exist. Literature Based Discovery (LBD) is the science of making more evident these connections. A particular problem that scientist face today is that of the evaluation of LBD system, which is very complex. In this article we consider some of the most important aspects that should be taken into account when evaluating an LBD system. We also present a preliminary exploration of the concept *of* relevance in this context.

Keywords: probabilistic topic models, evaluation, literature based discovery

1. INTRODUCTION

In the mid-80s Swanson encountered two disjoint literature bodies that were complementary i.e. when put together, they suggested an answer to a question which was not previously published. Swanson saw the potential in this procedure -- combining knowledge from both literatures to form an answer -- in generating new hypotheses, and started to systematically investigate this phenomenon under the name of Undiscovered Public Knowledge (UPK), more recently known as Literature Based Discovery (LBD)[13]. As a consequence of the natural fragmentation of science into specialities, disjoint but related literature does exist, some of Swansons' discoveries can be found in [12,14]. Therefore, new discoveries are waiting to be realised. Although the problems of fragmentation and specialisation in the context of literature search have been actively discussed in the past, this problem has not been studied extensively[15]. Research into LBD has started to address these issues.

For example, librarians are in a unique position to contribute to the development of science as information passes through their hands every day. Ashworth suggested that *''this need for combination is so fundamental that library systems should be designed to meet the requirement, thus enabling librarians to play an active and vital part in future innovation"*[1]. In Ashworth's proposed design we find that, amongst others, systems should have the ability to (i)demonstrate automatically when ideas are neighbours of each other, and therefore related, and (ii) uncover, automatically, valid statistical correlations between apparently unrelated matters. These proposals encapsulate the spirit of LBD research.

In this article we introduce further the problem of LBD (Section 2) before discussing the specific problem of evaluating LBD techniques (Section 3). As such, it is more of a position article which presents initial ideas and comments instead of solutions and results. However, we attempt to provide some conclusions and future work directions from this discussion (Section 4).

2. LITERATURE BASED DISCOVERY (LBD)

Swanson's observation was that science organises itself into manageable units and, while many of them are logically related, since they are created somewhat independently their relationships may be hidden even to their creators[12]. However, as Swanson illustrated, the creative use of online information seeking and retrieval applications could lead to the discovery and detection of potentially unintended logical connections by bridging the gaps between isolated literatures, resulting ultimately in new breakthroughs in science. This is the problem of LBD.
In its most basic form LBD could be presented as follows. Suppose that users are interested in finding the relationships between two concepts[1] A and C. Suppose also that these relationships are not discussed explicitly, i.e. there are no published documents discussing the relationship, or connection, between A and C. A boolean search for *A and C* would yield no relevant results[2]. It could be the case that A is related to an unknown concept X and that that same concept X is also related to C. Then, it could be argued that A is related to C through X. This

1 Almost all of the work done on LBD discusses concepts and connections between them. We will use the words concepts and topics interchangeably throughout the paper

2 For the LBD process to be justifiable, a search for *''A and C"* must yield no, or very few, relevant results.

implies that conducting two searches, *A and X* and *X and C*, could help the situation. However, since the concept X is unknown to the users, an exhaustive search for every possible concept X is certainly not the right technique. Users, in respect to this, are actually looking for those Xs such that the relationship between A and C is more manifest. The need for a system suggesting plausible X concepts, as argued by Ashworth and Swanson, relating A and C becomes more evident.

Models in LBD There are two major models in LBD as defined by Weeber et al.[18]. The *open model* is an exploratory model where users begin a literature search on a known concept A. From the resulting documents, a list of B concepts is extracted. This list is usually long so a post-processing step is needed where filtering/ranking is typically performed. The same procedure is then applied to every B concept and a list of new C concepts is presented to the user. The *closed model* is where users assume that a relationship exists between two known concepts A and C. Then, to validate this hypothesis, first a search for literature on ‘‘A or C’’ is done, then B concepts related to both A and C are extracted. Again, a post-processing step is taken on the resulting list.

LBD techniques adhere to one (or both) of these models, making the differences between them lie in how concepts are modelled and what techniques are used to post-process them. Initial attempts modelled concepts as words/phrases paying special attention to those appearing in the titles of the articles[16]. This raises the issue that most work has relied, up to some extent, on structured data making free-form text approaches rare. Other models of concepts have been evaluated such as clusters of documents and MetaMap concepts. The way concepts are modelled is related to the way they are acted upon. At the post-processing level, concepts are filtered, ranked, matched against users’ interests, etc. The techniques that have been applied at this level range from shallow term statistics such as term/document frequency[8] and higher level statistics such as LSI[4] to semantic filters and association rules[7]. For a complete review of these techniques the reader is encouraged to refer to [9].

A brief note on the models As Doyle pointed out *‘‘Literature searchers value both the unexpected and the expected. When something unexpected is found, one thereby obtains information; when the expected is found, one obtains confirmation. However, when one formulates a search, the unexpected is hardly ever involved. Search requests are practically always constructed out of familiar combinations of terms.”*[2].

Both models of LBD share a direct relationship with Doyle’s observation. By using the open model of search, a user will be confronted with the unexpected, with what is unknown to him yet related to the initial search. By using the open model, users should obtain information about the possibly related topics. As a complementary step, users could perform a search using the closed model. As a starting point, both the initial topic used in the previous search and a found topic could be used. Results from this search are actually evidence supporting different types of relationships between the initial topics. At this stage users obtain confirmation (or refutation) on something they suspected (or had found in the previous steps).

Although in the search of new knowledge the open model plays a major role, we believe that the process should be a two step process whereby one first discovers and chooses possible relationships of interest and then one searches for evidence supporting such relationships. The main point being made here is that the discovery of new connections (through exploration) and their validation (through the search for more evidence) are equally important to LBD evaluation.

Motivation A main aim of our research into LBD is to model how connections are discovered through the identification of intermediary concepts which form the bridge between disjoint literatures. To do so, we consider the use of probabilistic topic models (PTM)[11] for LBD. It has been argued that the Bayesian framework for probabilistic inference provides a general approach for understanding how problems of induction can be modelled in principle[11]. A Bayesian framework allows us to update beliefs in the light of new evidence based on assumptions of the problem in question and prior knowledge.
We will leave a thorough treatment on this motivation for PTM to future work as the main theme of this paper is the question of how to evaluate LBD. In the following section we wish to engage a discussion on the aspects we consider important for evaluation, as well as attempt to provide some hints of what the answers to these questions might be.

3. EVALUATING LITERATURE BASED DISCOVERY

The main purpose in evaluation, whether in Information Retrieval (IR) or any other area, is to measure the sensitivity of the measured variables to changes in system parameters. The evaluation methodology in Laboratory IR is fairly established, using test collections and relevance judgements as artefacts to simulate and measure these variables[17]. Relevance is considered to be a relation between a document and a query (where the query is an approximation to the user’s information need). Usually, this means that an expert judges the documents in the collection, regarding to the query, to be either relevant or non-relevant. Relevance, in this context, is interpreted as

'topicality', where 'topicality' means 'on topic'. This is useful to measure the performance at a system level, however this approach leaves several questions unanswered that, in the case of LBD, might be central to measuring the success of a system.

Previous analysis of LBD suggests that evaluating a system is non-trivial and may not be as fully defined as the paradigm adopted in IR. Almost all approaches to evaluate a system involved replicating Swanson's discoveries[12,14] and then observing how high the system ranked these already known connections. Since all of Swanson's discoveries were from discovering implicit connections in literature from the MedLine collection (a collection of medical scientific papers), in order to evaluate the performance of a system on different collections, authors have to resort to conducting user studies[3].

These studies also concentrate on evaluating a single aspect of the process (e.g. the generation of novel hypotheses). However, we believe that studying other aspects of LBD would be advantageous, especially on the user side, e.g. the learning process that might be involved, the (new) information a connection might provide, etc. This suggests that relevance cannot be taken to be mere topicality, as in laboratory IR, but to be taken as a multi-faceted entity including novelty, scope, belief and possibly more. We therefore keep in mind that such user studies make the comparison across systems subjective, but replicating Swanson's findings only will limit the evaluation of LBD systems to how successful they were at reproducing *those* findings and nothing else. Therefore, a more general mechanism of evaluation is required that will enable the extrapolation of LBD to scientific fields other than medicine. We now investigate this issue further.

A relevance model for LBD Psychological relevance is a theory of relevance for IR introduced by Harter in [6]. In Harter's theoretical model, a user's information need is actually taken to refer to the user's context or set of assumptions and facts known about the world. Therefore, relevance is modelled as a relation between a premise (a piece of information) and the current user's context. A document (or any other piece of information) is said to be psychologically relevant if and only if it produces a cognitive change in the user's context. A cognitive change can take place in different ways, such as the generation of new knowledge by the combination of previous assumptions and facts in the user's context with the new piece of information, old assumptions becoming stronger or weaker by the manifestation of new assumptions, etc.

When interpreted in LBD, a cognitive could mean that the user's context is modified by the addition of a new hypothesis about the relationship between the topics involved. If we accept this model of relevance, we can see how relationships suggested by an LBD system can be judged relevant or not. Given a suggested relationship, identified by the triplet (A,B,C), three possible outcomes can occur: the relationship is unknown to the user, the relationship is known to the user (possibly adding extra information) or the relationship is contradictory, i.e. it contradicts previous assumptions. If the relationship is unknown, the user will find it relevant. This is the best possible outcome and the ultimate goal of an LBD system. Unless the triplet conveys extra information, if it is known it will not cause a cognitive change and it will not be considered to be relevant to the user. Still exceptions can be observed. As Harter correctly points out[6], a known piece of information might be buried in the user's context to the point that it can be considered to be not known at the time the search is conducted. In this case, a known triplet acts as a reminder of something the user knows already. This triplet is considered relevant since it produces a cognitive change on the user's context. If the triplet does provide extra information, then it might be considered relevant as well. Imagine a triplet (A,B,C) where the relationship (A,C) is already know to the user, but through a different intermediate concept B', i.e. the triplet (A,B',C) is known to the user. In this case, the new suggested relationship is really a new aspect of the already known relationship between concepts A and C. This triplet is then considered relevant. The remaining outcome is also not an obvious one because a contradiction may or may not cause a cognitive change. If the evidence contradicting previous knowledge is strong enough for the user to believe it, a cognitive change will occur and the user's context will have changed (the previous piece of knowledge is replaced with the new hypothesis). In this case, a contradictory triplet (A,B,C) is considered relevant. If, on the other hand, the contradicting evidence is weak, the user's context will not change and the triplet is considered non relevant.

We argue that this model of relevance can accommodate the type of relevance sought after in an LBD system, a relevance model where relevant pieces of information, i.e. the connections between topics, generate these cognitive changes in the user's context and where the generation of new hypotheses is one of them.

New hypotheses: new to whom? A system implementing any of the LBD models should, by definition, suggest relationships that might end in the discovery of new knowledge. The question that may remain is *new to whom?*

Swanson's aims were to find novel hypotheses that would explain a hidden logical relationship between two apparently unrelated literature bodies. However it could be argued that the novelty of a hypothesis is relative to the user's context. One might be tempted to ask oneself to whom something qualifies as novel? To a five year old

child, a well known concept such as *sum* is unknown until it is explained to him. Gordon and Lindsay[3] relax the novelty restriction suggesting that the novelty in a connection might only be there to a particular user but not for a community. Does this mean that the connection lacks merit? If a connection is already known to the community, not all is lost. The connection might still be new or interesting to the user operating the system. In this respect, LBD can be seen not just as a hypotheses generation/evaluation model but also as a model to aid learning, e.g. think of a scientist migrating from a field to another one. The transition can be made easier if the commonalities between areas (the researcher's original area and the new area) are made more evident by such a system.

In Harter's theory of relevance, a fact is manifest to a user only if he is able to understand it and accept its representation as true or probably true. This suggest that the user's previous expertise in an area, which is a part of his context, is important when assessing the relevance of a piece of information, which in LBD is represented as a connection[10]. To exemplify this situation suppose an experiment where two sets of users with different contexts, e.g. a set where users are experts in an area and a set where users are students in the same area, are asked to evaluate the relevance of the results presented by a system, i.e. the connections found. Clearly, users in the group of students may find certain connections hard to interpret or to be non relevant because they lack the previous necessary knowledge. At the same time, a user from the group of experts may find this same connection as relevant and develop a hypothesis explaining the relationship between the topics.

The novelty of the information presented, as part of its relevance, is relative to the user's context, much in accordance with the theory of psychological relevance, and it should be measured at different levels or scopes.

Rejecting coincidences and PTMs How do we know that a particular triplet (A,B,C), even though highly correlated, is not just a coincidence? From a system's point of view, all that can be assured is that a particular topic B is highly correlated to both A and C and that some other properties are met (reflected in the scoring function of the triplet).

Griffiths et al.[5] suggest that coincidences can be best interpreted in a framework of rational statistic inference. Given a set of hypotheses *H*, where each member of this set is a theory of how data is generated, a learner can assign prior probabilities *P(h)* to each *h in H*. After observing some data *d*, using Bayes' rule we can calculate the posterior probability of a particular hypothesis *h* as $P(h|d)=P(d|h)P(h)/P(d)$ where the likelihood $P(d|h)$ is the probability that *d* is actually generated by the *hypothesis h.* Using Griffiths' words then, "*coincidences arise when there's a conflict between the evidence an event provides for a theory and our prior beliefs about the plausibility of that theory"[5]*.

In LBD, the case where co-occurrences between A and an intermediate topic B and C and the same intermediate topic B are mere coincidences is not something unrealistic to observe. In fact, since the number of connections that can be proposed, if based only on co-occurrence of words, will be rather large, the chances that most of them are meaningless to an individual at a given time[3] are high. This is a well known problem and the way different systems deal with it is by ranking the list of intermediate topics according to what the system believes are potentially more relevant. On top of ranking, or as a previous step, some systems also filter out the suggestions that are believed to be noise rather than potentially useful information. Both operations are applied based on the score of a suggested relationship where the score is calculated using different sources of evidence.

Our approach at ranking the suggested relationships by a system is based on this theory of coincidences. We wish to reject as much as possible coincidences, therefore favouring suggestions for which there is not only strong evidence but also a comparable prior belief. This leads to a more formal and intuitive model for ranking the proposed relationships. Our proposed model of LBD makes use of probabilistic topic models (PTM)[11] to model topics and instantiate Griffiths' model of coincidences.

Exploratory search Traditional IR systems are systems that answer questions of the *what* type, providing users with information on the topic provided. On the other hand, the goal of an LBD system is not to answer questions but rather to provide evidence supporting a particular hypothesis (closed model) or to provoke thought on the user's side (open model) to help him generate a hypothesis, i.e. *there is some evidence that A might be related to C through X, how could A, X and C interact together?* Moreover, LBD systems are designed to be a supporting tool in the process of hidden public knowledge discovery. In that sense, a desirable property in a system is that it covers as much of the possible connections space as possible. Given two topics, A and C, there might be more than a single possible valid connection, therefore a system suggesting more valid possible connections should be considered more helpful than systems suggesting fewer.

3 That is, non relevant

In an era of information explosion, there are not many practical uses for recall-oriented systems, yet LBD may be one of them. To make sure that a particular hypothesis is new, a searcher needs to explore the information space as much as possible. Some authors have referred to this as *negative search*, where a user actually expects a search not to return any relevant documents. Complementary to this is the situation when a user is considering the possible relationships between two concepts. If a system discards the only one that leads to a valid discovery then that system should be considered a failure. In respect to this, think of legal search where failing to retrieve a particular relevant document puts the entire prosecutor's case at risk.

Recall is a very important aspect in LBD since the coverage of the available information can easily establish whether a theory can be considered new. At the same time, measuring Precision in LBD might not even make sense since assuming that a document—or a proposed relationship—is automatically non relevant may lead to unrealistic results (when it hasn't been judged either relevant or non relevant).

4. CONCLUSIONS AND FUTURE WORK

More and more we see examples where the interpretation of relevance as mere topicality is not enough and perhaps a too shallow commitment. In LBD, especially, topicality is not what a searcher is after but something almost tangential to his search. We have argued that a more comprehensive model for relevance is needed in LBD and that perhaps Harter's model suits the purpose. Using Harter's definition of relevance, we see that a relevant suggestion changes the user's context in such a way that a new hypothesis is generated. However, relevance as a multi-faceted entity implies that the evaluation of a system is non-trivial and, in order for any evaluation to be meaningful, several factors have to be discussed first. We have suggested some aspects which we think play an important role such as novelty, coverage and the rejection of obvious dead ends. Still we are far from having a well defined methodology for evaluation. Hopefully this article will serve its purpose, to take one of those small, but needed, initial steps towards an answer to this question.

REFERENCES

[1] W. Ashworth. Librarianship and other disciplines. Midlands Branch A.G.M., 1966.
[2] L.B. Doyle. Semantic road maps for literature searchers. Journal of the ACM, 8(4):553–578, 1961.
[3] M. Gordon, R.K. Lindsay, and W. Fan. Literature-based discovery on the world wide web. ACM Transactions on Internet Technology, 2(4):261–275, 2002.
[4] M.D. Gordon and S. Dumais. Using latent semantic indexing for literature based discovery. Journal of the American Society for Information Science, 49(8):674–685, 1998.
[5] T.L. Griffiths and J.B. Tenenbaum. From mere coincidences to meaningful discoveries. Cognition, 2006.
[6] S.P. Harter. Psychological relevance and information science. Journal of the American Society for Information Science, 43(9):602–615, 1992.
[7] D. Hristovski, B. Peterlin, J.A. Mitchell, and S.M. Humphrey. Improving literature based discovery support by genetic knowledge integration. Studies in Health Technology Informatics, 95:68– 73, 2003.
[8] R.K. Lindsay and M.D. Gordon. Literature-based discovery by lexical statistics. Journal of the American Society for Information Science, 50(7):574–587, 1999.
[9] C. Murat, M. Pottenger, and C. D. Janneck. Recent advances in literature based discovery. Technical report, Lehih University, 2005.
[10] I. Ruthven, M. Baillie, and D. Elsweiler. The relative effects of knowledge, interest and confidence in assessing relevance. Journal of Documentation, 2007.
[11] M. Steyvers and T.L. Griffiths. Probabilistic topic models. Latent Semantic Analysis: A road to meaning, 2005.
[12] D.R. Swanson. Fish oil, raynaud's syndrome and undiscovered public knowlege. Perspectives in Biology and Medicine, 30:7–18, 1986.
[13] D.R. Swanson. Undiscovered public knowledge. The Library quarterly(Chicago, IL), 56(2):103–118, 1986.
[14] D.R. Swanson. Migraine and magnesium: eleven neglected connections. Perspectives in Biology and Medicine, 31:526–557, 1988.
[15] D.R. Swanson. Asist award of merit aceptance speech. ASIST, 2001.
[16] D.R. Swanson and N.R. Smalheiser. An Interactive System for Finding Complementary Literatures: A Stimulus to Scientific Discovery. Artificial Intelligence, 91(2):183–203, 1997.
[17] E.M. Voorhees. The philosophy of information retrieval evaluation. Evaluation of Cross- Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF, 2406:9–26, 2001.
[18] M. Weeber, H. Klein, L.TW de Jong-van den Berg, and R. Vos. Using Concepts in Literature-Based Discovery: Simulating Swansons Raynaud–Fish Oil and Migraine–Magnesium Discoveries. JASIST, 52(7):548–557, 2001.