Identifying Scientific Breakthroughs by Combining Co-citation Analysis and Citation Context

Henry Small[1] and Richard Klavans[2]

[1] *hsmall@mapofscience.com*
SciTech Strategies, Inc., Bala Cynwyd, PA 19004 USA

[2] *rklavans@mapofscience.com*
SciTech Strategies, Inc., Berwyn, PA 19312 USA

Abstract

This study combines two relatively independent approaches to identifying scientific breakthroughs from an analysis of the scientific literature. The first approach focuses on citation network data that is gleaned from the bibliography of scientific documents. The second approach focuses on the text in these documents (especially the text surrounding a reference, the so-called citation context). We have linked these two approaches in order to identify potential scientific breakthroughs within three research communities. The analysis is facilitated by creating a large scale categorization of words and phrases called modalities that scientists use to enhance or diminish the credibility of scientific statements. The rates of occurrence of the modality categories for research communities, papers and sub-regions of those communities are examined to provide indicators for changes in research direction.

Introduction

There are two general approaches to forecasting scientific breakthroughs using scientometrics. The first, and more popular approach, is limited to the analysis of the references at the end of scientific documents. This research stream, greatly facilitated by the creation of the ISI databases in the 1970's, has tended to focus on citation frequency. Basically, highly cited references that were recently published are considered breakthroughs. Clusters of highly cited references indicate that there is a concentration of breakthroughs that deserve special attention (Upham & Small, 2010).[21] Chen and co-workers have provided a comprehensive theoretical and empirical model for this approach (2009).

The second approach is based on an early criticism of citation frequency (Moravcsik & Murugesan, 1975). Researchers pointed out that no adjustments were made for the context of the citation. For example, a paper might be highly cited because it was wrong. Or a paper might be highly cited because it is symbolic of a broad area of research. The second approach (citation context) therefore focuses on the text surrounding a reference.

These two approaches have remained relatively independent because of data availability. The citationists lean toward using databases that represent 'all of science' (about 1 million articles per year). None of these databases provide full text data, thereby precluding any citation-context analysis on a large scale. The contextualists lean towards using relatively homogeneous (and small) sets of scientific documents where there is full text. This approach is not easily generalizable across many areas of science and is subject to selection bias (what documents are or are not included in the target literature).

These two approaches have been integrated using the Scopus database (about 1 million articles per year) and full text (for about 300,000 articles per year). This is the first time that researchers have attempted to integrate these two approaches at this scale of analysis. The following describes how this integration can proceed and provides three concrete examples.

[21] Co-citation analysis is one of the most common methods for identifying clusters of highly cited references.

Rationale

To understand how citation context may enhance our ability to predict new directions in science, we consider how references occur in their native setting, the scientific text. First at a physical level, the text provides information on the sequencing of references in the paper, their distribution over sections, paragraphs, and sentences. This distribution is often very clumpy, with more references occurring in certain sections or paragraphs than others, and with some references repeated multiple times in the text and others cited together with other references at a specific text location (Maricic, et al., 1998). At the lexical level, each reference point is embedded in a text segment, and in the contextualist's approach this can be mined for terms bearing directly or indirectly on the reference. If patterns or categories of such words can be associated with a particular cited reference, for example, words indicating that the item in question is new, novel, or important in some way, then this information could augment citation counts or dates of publication, to enhance our ability to detect breakthroughs or shifts in research direction.

By extension, when a text cites two references, the classical definition of co-citation, the two references can be in close textual proximity or far apart. When they are in close proximity, which is commonly the case for highly co-cited items (Small, 2011), there is the likelihood that citation contexts will be shared. This means that breakthrough-indicating words will tend to associate with clusters of documents, and breakthrough designations can be applied at this level as well.

Citation Context

Citation contexts have seen varied applications over the years. Small (1982) differentiated two types of studies: the classification of the function or motivation of references in the citing text, and the use of the semantic content of contexts to characterize specific cited works. Lately there seems to have been a minor resurgence of interest in this area, particularly among scholars in computer science and linguistics. Some of these studies have used citation contexts as a way to automatically add indexing terms to the cited papers under the assumption that citation contexts provide thumbnail descriptions of cited papers and can improve retrieval (Ritchie, 2008; Elkiss et al., 2008). Other studies use citation contexts to label the nature or function of citations in scientific text by means of language based algorithms to automatically extract and classify contexts (Teufel, 2010). Another recent theme is to use what have been called citation sentiments to label and interpret regions or links on a map of science and relate categories to structural characteristics of maps such as inter- or intra-disciplinarity.

There have been numerous proposals over the years on how citations should be classified. These efforts have often resulted in similar categories across schemes, despite differences in category names. The reason for this is in part pragmatic. Scientists provide only limited language cues to judge intention and motivation in the highly conventionalized style of scientific writing (Swales, 1990), and analysts have therefore responded to what cues are available to construct categories.

Many earlier schemes have focused on citation function, that is, what function the reference serves for the citing author. This approach can entail trying to infer an author's intentions or motives which is difficult given available cues. However, little attention is given to how the context reflects on the process of scientific research - whether it involves a discovery, hypothesis, or an analogy - or the status of the knowledge under discussion - whether it is certain or uncertain, important, or new. In 1979 Latour and Woolgar introduced the notion of knowledge modalities that are used by scientific authors to modify the status of scientific statements. In their theory, words such as "reported", "first", "convincing", "difficult", "support" and "suggested" are used to modify the degree of tentativeness or certainty in the underlying knowledge. Only when such modalities are omitted can the statement be considered to be tacit, taken for granted, and a scientific fact.

We propose to use Latour's model as a basis for a citation classification. The approach taken integrates research process factors by relating categories to an epistemological scale, on the one end, from knowledge that is considered incorrect or uncertain to the other end, where knowledge is generally accepted or taken for granted (see also Finney, 1979). A similar scale was, in fact, used by Latour (1979, p. 82; 1987, p.44) to show stages in the evolution of a scientific fact, and thus the scheme attempts to capture aspects of the process and progress of scientific knowledge.

Table 1 lists the categories ordered roughly on the basis of the certainty of the knowledge. For example, "causality" is near the top of the scale because discussion of causes usually implies knowledge that is generally accepted. "Similarity" is near the bottom because inference on the basis of similarity or analogy is risky. The category "constructed", defined as dealing with the fabrication or composition of objects, is near the middle of the scale because devices can embody known principles but their operation may not be completely understood. Moving up from the bottom of the scale there is a progression of research from "uncertainty" to "new" to "hypothesis" to "modified" to "importance", a plausible scenario for a progressive research program or stages in problem solving.

Table 1. Knowledge modality categories ordered from most to least certain.

Order	Category Name	Order	Category Name
1	Established	14	Method
2	Previous	15	Usage
3	Causality	16	Interest
4	Consensus	17	New
5	Importance	18	Differentiated
6	Discovered	19	Associated
7	Achieved	20	Similarity
8	Improved	21	Difficulty
9	Modified	22	Future
10	Supported	23	Uncertainty
11	Hypothesis	24	Weakness
12	Reported	25	Criticism
13	Constructed	26	Negation

In contrast with earlier schemes, the categorization proposed here is concerned with how authors characterize ideas represented by cited work, and not necessarily the cited work itself, and hence, while designed using citation contexts, could in principle be applied to any scientific text. Despite this difference in approach, the proposed scheme does overlap significantly with one proposed by Garzone and Mercer (2000) who end up with 35 categories as an amalgam of previous proposals. This demonstrates the previous point that citation categorization researchers have only limited cues to work with.

There have also been differing approaches to how citation classifications are implemented. Early work relied on human judgment to assign contexts to categories. Later categories were associated with cue words which were matched against contexts (Finney, 1979). Most recently Teufel (2010) has employed a much more detailed linguistic parsing of the full text of each article, utilizing a complex set of word patterns and formulaic expressions which then are processed through a machine learning technique. Teufel classifies each context to only one category which seems like a severe limitation. The cue word-modality scheme proposed here assigns a given context to multiple categories, and provides a spectrum on the certainty scale.

After categories were decided on, the next step was to build an extended lexicon of modality words and phrases assigned to each of the categories. The lexicon was built up by a combination of methods. First, random samples of contexts from three communities were drawn and manually categorized, noting the words and phrases that prompted the categorization. Second, word counts were created for the full sets of contexts and manually scanned for words relating to the categories. This process was facilitated by screening out technical vocabulary by use of a common word list. Thirdly, lexicons from prior studies (Small, 2011; Teufel, 2010; Nanba, Kando & Okumura, 2000), were scanned for additional terms, and to a limited extent synonyms were drawn from resources such as Wordnet (Miller, 1995). The lexicon used for this study consisted of about 1000 words and phrases, including grammatical variant forms.

The main challenge with the cue word/phrase approach is to avoid matching on incorrect word forms. Because modality words can also mimic technical terms, it can be difficult to separate the two usages. Queries need to be carefully tuned to minimize retrieval of non-relevant contexts. For example, in the brain imaging community, the word “attention” is used as a technical term, while in other fields it is used to denote “interest”. Of course, modality words can reference either the cited work in a context or some object of research, such as “important work on estrogen” and “the important hormone estrogen”.

Integration of Co-Citation Analysis and Citation Context

A 2007 model of science, using co-citation analysis, was developed from the Scopus database using the STS methodology (Boyack & Klavans, 2010). This resulted in a total of 92,097 research communities. Each community consisted of a set of current citing papers from 2007 and a set of base papers that are cited by the current papers. There were 2,317,808 unique references and 1,443,005 current papers in this model of science.

Full text data, for 2007, was also made available from Elsevier. Elsevier is the largest single publisher of scientific literature, and as such, is most likely to have relatively high coverage of the current papers in a large number of research communities. These data were matched to the 2007 science model, resulting in a relatively large number of research communities with high full text coverage.

For this exploratory study, we focused on three relatively large research communities from different fields of science that had high full text coverage. The selected communities were water pollution by hormonal-type substances in environmental science, the development of biosensors in materials science, and brain imaging studies in neuroscience. The availability of full text for a significant fraction of the 2007 citing papers enabled us to study the ways in which the base papers for these communities were cited in 2007.

To extract the citation contexts for analysis, the bibliographic data from the 2007 model had to be matched with the full text data. First the reference and citation contexts were extracted from the full text. Because the XML format anchors the point of reference within the text to the reference at the end of the paper, it was possible to automatically connect the citation contexts in each citing paper with the specific cited document in the reference list. Then the reference information from the full text was matched against Scopus data to attach Scopus identifiers to the cited references, and also to the 2007 citing papers. Finally the model’s base papers were matched against the Scopus identifiers.

The match rate of the 2007 model’s references was around 68%, meaning that over two-thirds of the references cited by the full text papers could be assigned to an STS research community, together with their associated citation contexts. Also for each full text paper, all references were included whether or not they were assigned to a community in the 2007 model and regardless of what community they were assigned to. This provides citation contexts for references that are made to base papers within a community as well as context for references outside the community. For the three communities in this study, there is a ratio of about one endogenous citation for every four exogenous citations. Table 2 gives the size of the communities in terms of base and citing papers, the percentage of full text coverage, the number of distinct contexts, and average context lengths.

Table 2. Statistics on the selected research communities.

Research Community	base papers	citing papers	full text coverage	citation contexts	mean context length
Hormone pollution	72	192	46.3%	3,134	48.4 words
Biosensors	90	181	47.0%	3,014	47.8
Brain imaging	79	189	39.1%	3,332	51.1

Citation Contexts

A citation context was defined as up to three sentences around the reference indicator in the citing text: the sentence the reference occurs in, sometimes called the "citance" (Nakov, Schwartz & Hearst, 2004), the sentence preceding the "citance", and the sentence following the "citance", provided that the preceding or following sentences do not contain another reference and do not cross a paragraph boundary (Nanba, Kando & Okumura, 2000). This definition resulted in contexts consisting on average of 1.9 sentences for the communities studied, which is comparable to 1.6 sentences found in a prior study where contexts were identified manually (Small, 2011). In the current data a closer examination revealed that contexts in one of the research communities, brain imaging, contained on average 6% more words than contexts in the other two communities.

Access to the full text of papers and their citation contexts allows one to compute new metrics which are not easily accessible using standard citation indexes. For example, we can compute what might be called the *op. cit.* rate, the number of times individual references are cited within a given citing paper. Across the three communities the *op. cit.* rate is 1.5. The second metric is the rate at which references in a paper are cited in bunches at specific points in the paper. This might be called the redundancy rate, following Moravcsik and Murugesan (1975) who first noted the phenomenon. The overall redundancy rate for the three communities is 1.6, meaning that 1.6 references on average are cited at a given reference point. Of course this means that the citation contexts for the set of papers cited redundantly will be identical. Access to *op. cit.* and redundant citations at the intra-textual level opens up the possibility of new types of citation and co-citation counting (Teufel, p. 63). They can also enrich citation and co-citation context studies because, for example, a citing paper could yield multiple citation and co-citation contexts for a single cited document, and redundancy can be used to measure document equivalencies.

Results

Modality Comparisons

Our main research objective is to determine whether modality categories occur at different rates in different research communities and subsets of citation contexts, and whether such match rates could be indicative of scientific breakthroughs. The procedure is to match the word and phrases for each category against different sets of contexts counting the number of matches per category. The fraction of contexts matching a category for each sample can then be compared, for example against a baseline, to determine statistical significance.

The first result is a match of the lexicon against all contexts for each of the communities. A count was made of the number of matching contexts for each modality category. The match rate for the three communities combined was used as the baseline for a calculation of significance using the log likelihood statistic.

Table 3 reports the top two modalities for each community by log likelihood, and the number of categories with significant match rates at the p < .01 level. The high number of categories for brain imaging may be in part due to the longer average length of the contexts for this area (see Table 1), a tendency toward verbosity, or more frequent use of modality words. These alternatives can be tested in the future using a word level match rate, rather than a context count rate, which would effectively normalize for context length.

Table 3. Top two knowledge modalities for each research community, and total number of significant modalities at p < .01.

Research Community	Hormone pollution	Biosensors	Brain imaging
Modality 1	Causality	Usage	Associated
Modality 2	Weakness	Constructed	Hypothesis
Significant modalities	2	6	14

Despite this limitation, the results illustrate how modalities can interact with the community subject matter. The “causality” category for hormone pollution reflects the frequent attribution of adverse biological effects of estrogen-like contaminants in the water supply, while the “weakness” modality indicates some limitation in their methods. In biosensors the “construction” modality is the result of a focus on device fabrication. For brain imaging the “association” modality reflects research efforts to associate different areas of the brain with different mental states, and the “hypothesis” modality suggests the importance of theory making.

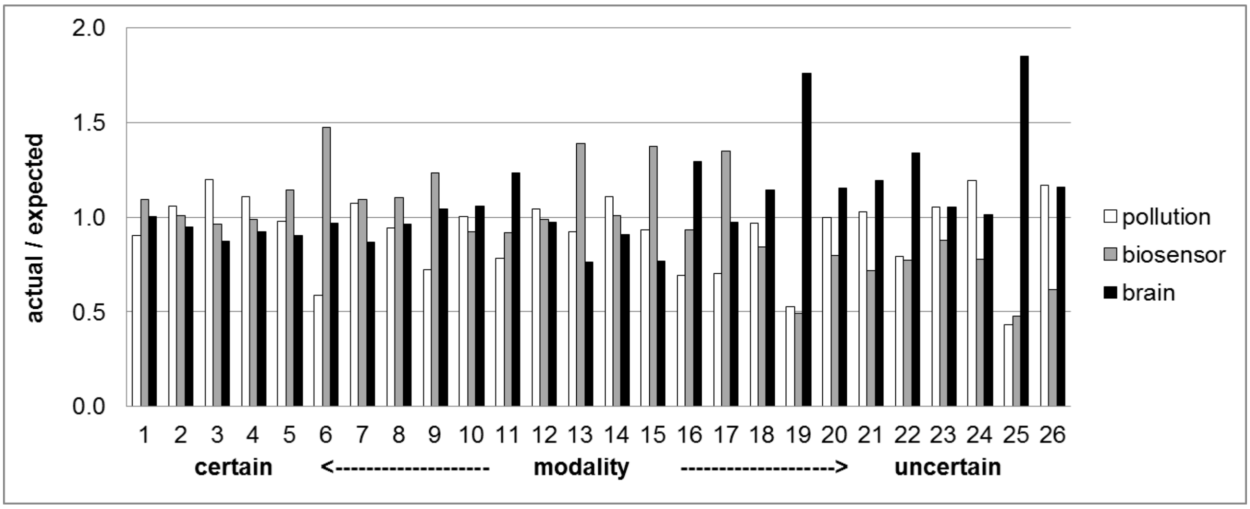

Figure 1. Modality spectrum for three research communities. Modalities are numbered 1 through 26 (see Table 1). The vertical scale is the actual number of contexts divided by the expected number based on marginal totals.

Another way of displaying this information is what might be called a modality spectrum (Figure 1), arranging the modalities from most to least certain. The actual to expected number of contexts for each modality is plotted on the vertical axis. Expected values are computed as in chi-square, based on marginal totals for each community and modality. The chart shows spikes for brain imaging on the "associated" modality, as noted above, but also on "criticism". The biosensor community has a strong spike for the "discovered" modality. The brain imaging community seems more heavily weighted toward the uncertain end of the scale, while the biosensor community tilts slightly toward certainty.

The second comparison looks at citation contexts internal to research communities as compared to contexts external to those communities, so-called endogenous versus exogenous citations. Earlier work had suggested that interdisciplinary or exogenous citations may express greater uncertainty (Small, 2011). Because our data sets included all references and contexts from papers citing a community, it was possible to examine this outside/inside effect. To obtain a set of endogenous citation contexts, citing papers were identified having more references to the base papers for the target community than any other community. It was assumed that any additional references made by this set would be exogenous because the other communities were being referenced less frequently. The two groups were then matched against the modality categories and results aggregated across the three communities.

Only one modality had a statistically significant over-representation in the exogenous set, namely "hypothesis" ($p < .01$). One interpretation is that the theoretical basis for these communities is coming from other areas. This is consistent with their somewhat applications oriented slant. Two other categories showed a significant endogenous representation, namely "associated" and "previous". The latter shows the importance of "previous research" within these fields.

Community Maps

We next turn to how modalities vary over the internal structure of the community. To examine this phenomenon, maps were created for the base papers in each community. Co-citations were computed for each pair of base papers and normalized using the cosine measure, taking the two strongest links per base paper. These data were input to Pajek software (De Nooy, Mrvar & Batagelj, 2005) using the Kamada-Kawai option. Figure 2 shows the map for hormone pollution. Five regions were selected having a high density of co-citation links for different random starting configurations in Pajek. Both modality and content word analyses were carried out for each region. Significant content words and phrases were selected using Wordsmith Tools software (Scott, 2010), and regions were labeled with the phrases containing the words with the highest log likelihood. Modality analysis was carried out by combining the contexts for all base papers in a region and matching against the lexicon. Selection of the most significant modality for each region was made using the contexts for the community as a whole as the baseline. The strongest modality for each region is given in parenthesis under the content label. Also given beneath the region label is an average certainty score computed by coding the nine modalities at the most certain end of the scale as +1 and the nine modalities at the uncertain end as -1.

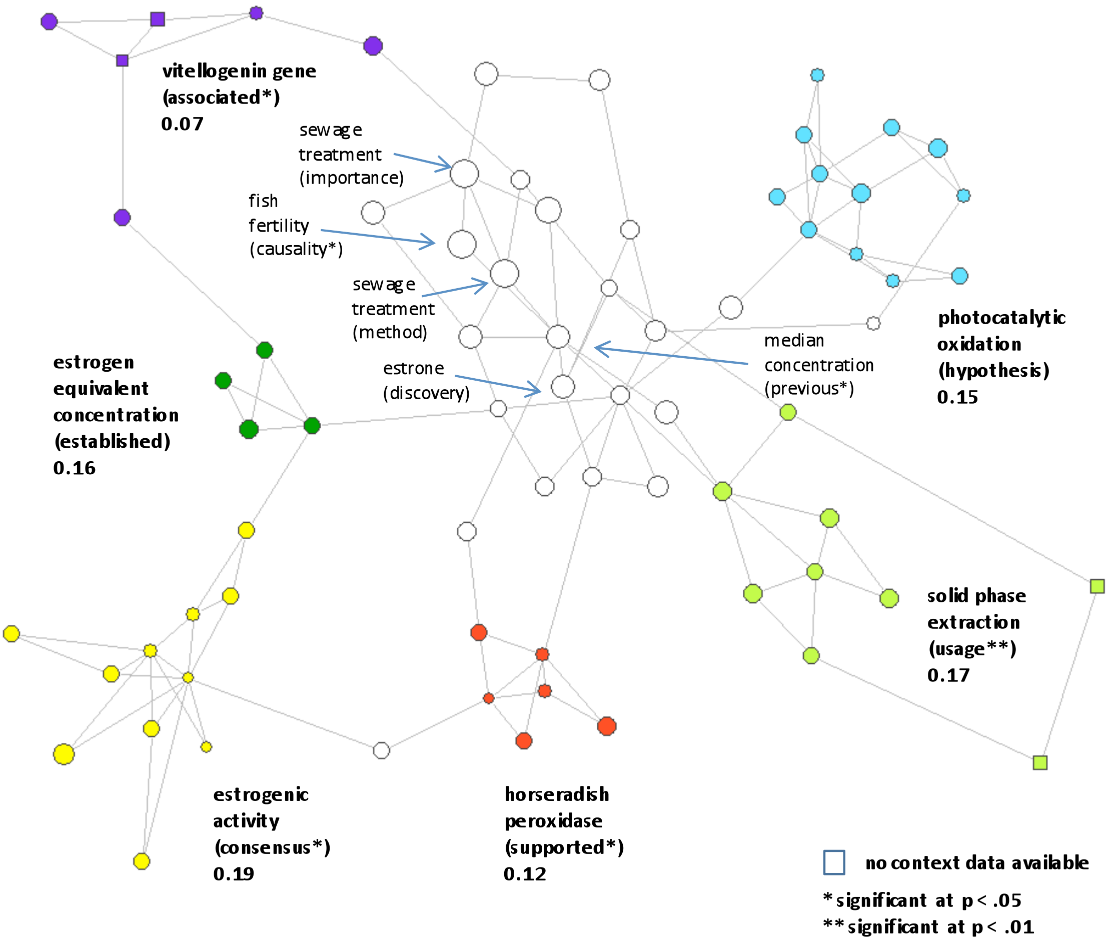

Figure 2. Map of the hormone pollution community based on co-citations. Regions papers are labeled with significant content words followed by modality in parentheses and mean certainty.

Even though the regions were identified by visual inspection, it is interesting to note that the context redundancy was 1.5 times greater for the regions than for the communities as a whole (2.3 citations per reference point versus 1.6). This means that the base papers that are closely tied by co-citation tend to have a higher concentration of redundantly cited papers, that is, papers cited at the same reference point in the citing papers, and therefore sharing the same context.

In addition to regions, five highly cited papers on each community map were selected for modality and content analysis in the manner described above. The papers are labeled with content words and arrows point to their location. One paper labeled "sewage treatment" is assigned the "importance" modality, and another labeled "estrone", the "discovery" modality. These may be predictors for the later evolution of the community. Other papers and regions, however, do not suggest an area on the verge of a major breakthrough, but rather one that is working within an established paradigm.

In most cases there is a direct tie between the technical label for a region or paper and the associated modality. For example, the region labeled “horseradish peroxidase” with the modality “supported” has the context: “A previous study *showed* that steroid estrogens . . . can be effectively oxidized by HRP [horseradish peroxidase] . . .” (italics added by authors). In this case the word “showed” is a cue word for the “supported” modality. For the region labeled “estrogenic activity” and “consensus” we have: “As in *various other* European countries, a nation-wide survey has been performed in the Netherlands . . . to make an inventory of the presence of estrogenic active substances in Dutch surface waters.” For the region labeled “vitellogenin gene” with the modality “associated” we have: “Exposure to potent estrogens . . . induced male VTG [vitellogenin gene] that was *associated* with impaired reproductive output in fathead minnow . . .” Finally for the paper labeled “fish fertility” with the “causality” modality we have the context: “Similar to the steroid hormones, pharmaceuticals as environmental contaminants did not receive a great deal of attention until the link was established between a synthetic birth-control pharmaceutical and *impacts* on fish.” Here the word “impacts” is a cue word for the “causality” modality. These examples illustrate how modalities are connected to technical issues, and we can expect breakthrough-indicating modalities to have similar connections.

Figure 3 for the brain imaging community shows again the prominence of the “association” modality for this topic. The modality of “new” for the region labeled “regional homogeneity method” suggests a possible growth point, as do the modalities of “new” and “future” for two individual papers.

Assessment

Another task for the future is a comprehensive assessment of how well the modality lexicon identifies the categories of the context in terms of recall and precision. A given context could match incorrectly on a modality, or fail to be categorized. A preliminary examination of 33 modalities that were significant at least at the $p < .05$ level for a combination of 19 map regions and highly cited papers, involving about 500 contexts, showed that five modality assignments were incorrectly assigned due to word ambiguities, giving a 16% type-1 error rate. This error rate is comparable with that obtained by Garzone and Mercer (2000). Errors were mainly due to words which had technical as well as modal meanings, such as “current” or “potential” in biosensors. Some of these matching problems could be overcome by part-of-speech tagging.

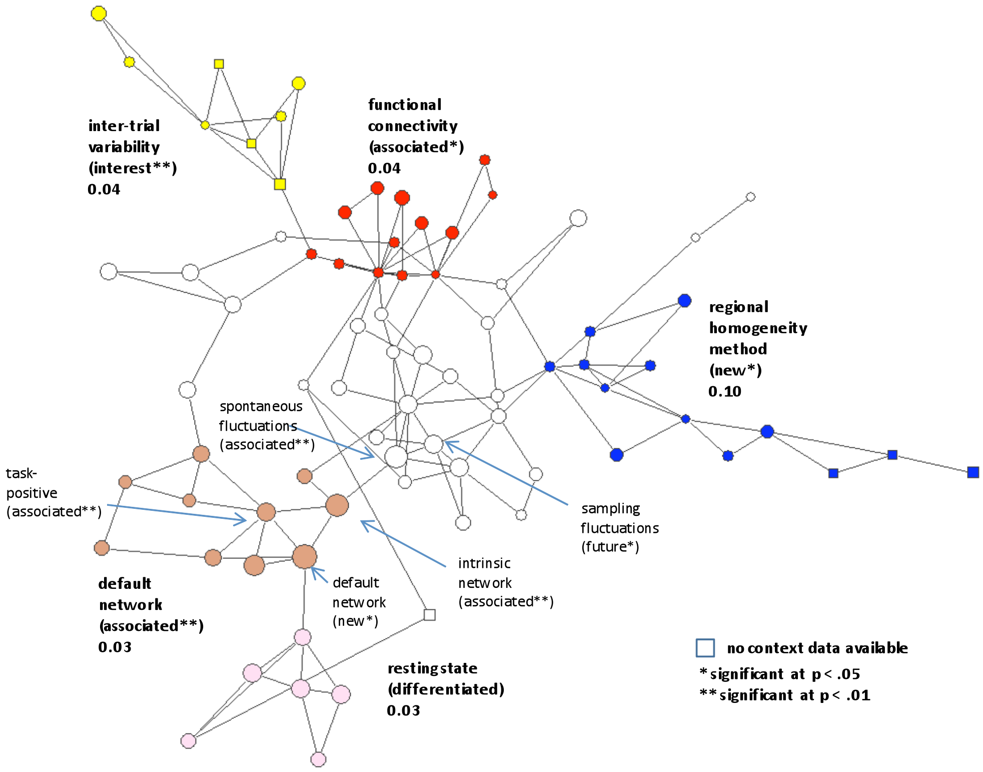

Figure 3. Map of the brain imaging community. Regions and papers are labeled with significant content words followed by modality in parentheses and mean certainty.

Conclusions

The availability of full text of scientific papers with XML mark-up of references opens up many new avenues for analysis. In particular, it appears feasible to apply citation context analysis to help in the prediction of breakthroughs and new directions in science. We have proposed a scheme based on a scale of certainty of scientific knowledge and cue words and phrases based on Latour's notion of modalities in scientific text. We have seen that modalities can be directly tied to research issues at the community level, such as specific problems addressed and methods of working. It remains to be seen whether these categories are predictive of change at the community or document level. Categories such as "importance", "discovery", "new" and "future" are clearly possible breakthrough indicators. Another issue is the location of a community or sub-region on the scale of certainty. Location at the high uncertainty end could signal pending improvement or imminent demise. It is also possible that the direction of shifts over time along the modality spectrum could be predictive of change. Progressive "moves" or problem shifts from low to high certainty are one possible indicator. It will be important to study communities over time to see if modalities are fixed or change in consistent ways, and whether they are correlated with changes in community size or citation rate.

Acknowledgments

Full text in XML for the publication year 2007 was generously provided by Elsevier under an agreement with SciTech Strategies, Inc.

References

Boyack, K.W. & Klavans, R. (2010). Co-citation analysis, bibliographic coupling and direct citation: which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology,* 61 (12), 2389-2404.

De Nooy, W., Mrvar, A. & Batagelj, V. (2005). *Exploratory Social Network Analysis with Pajek.* Cambridge: Cambridge University Press.

Chen, C., Chen, Y., Horowitz, M., Hou, H., Liu, Z. & Pelligrino, D. (2009). Towards an explanatory and computational theory of scientific discovery. *Journal of Informetrics,* 3, 191-209.

Elkiss, A., Shen, S., Fader, A., Erkan, G., States, D. & Radev, D. R. (2008). Blind men and elephants: what do citation summaries tell us about a research article? *Journal of the American Society for Information Science,59* (1), 51-62.

Finney, B. (1979). *The reference characteristics of scientific texts.* Unpublished Master's Thesis. The City University of London.

Garzone, M. & Mercer, R. E. (2000). Towards an automated citation classifier. In *Proceedings of the 13th Biennial Conference of the CSCI/SCEIO* (AI-2000) (pp. 337-346).

Latour, B. & Woolgar, S. (1979). *Laboratory Life: the social construction of scientific facts.* Beverly Hills, California: Sage.

Latour, B. (1987). *Science in Action*. Cambridge, Mass., Harvard University Press.

Miller, G.A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM*, 38 (11), 39-41. http://wordnet.princeton.edu.

Maricic, S., Spaventi, J., Pavicic, L. & Pifat-Mrzljak, G. (1998). Citation context versus frequency counts of citation histories. *Journal of the American Society for Information Science,* 49 (6), 530-540.

Moravcsik, M.J. & Murugesan, P. (1975). Some results on the function and quality of citations. *Social Studies of Science*, 5, 86-92.

Nakov, P.I., Schwartz, A.S., & Hearst, M.A. (2004). Citances: Citation sentences for semantic analysis of bioscience text. In SIGIR '04 Workshop on Search and Discovery in Bioinformatics.

Nanba, H.,Kando, N.,& Okumura, M. (2000). Classification of research papers using citation links and citation types: towards automatic review article generation. In American Society for Information Science SIG Classification Research Workshop: Classification for User Support and Learning.

Ritchie, A. (2008). *Citation context analysis for information retrieval*. Ph.D. thesis, Computer Laboratory, Cambridge University.

Scott, M. (2010). *WordSmith Tools Step by Step*. http://www.lexically.org/wordsmith/step_by_step_guide_english.pdf.

Small, H. (1982). Citation context analysis. In B. Dervin, & M.J. Voight (Eds.), *Progress in Communication Sciences,*vol.3, (pp. 287-310). Norwood, N.J.: Ablex Publishing Corp.

Small, H. (2011). Interpreting maps of science using citation context sentiments: a preliminary investigation. *Scientometrics*, in press.

Swales, J.M. (1990). *Genera analysis: English in academic and research settings.* Cambridge: Cambridge University Press.

Teufel, S. (2010). *The structure of scientific articles: applications to citation indexing and summarization.* Stanford, California: CSLI Publications.

Upham, S.P & Small, H. (2010). Emerging research fronts in science and technology: patterns of new knowledge development. *Scientometrics,* 83 (1), 15-38.