Delineating the Citation Impact of Scientific Discoveries

Chaomei Chen
Drexel University
3141 Chestnut Street
Philadelphia, PA
19104-2875
cc345@drexel.edu

Jian Zhang
Drexel University
3141 Chestnut Street
Philadelphia, PA
19104-2875
jz85@drexel.edu

Weizhong Zhu
Drexel University
3141 Chestnut Street
Philadelphia, PA
19104-2875
wz32@drexel.edu

Michael Vogeley
Drexel University
3141 Chestnut Street
Philadelphia, PA
19104-2875
msv23@drexel.edu

ABSTRACT

Identifying the significance of specific concepts in the diffusion of scientific knowledge is a challenging issue concerning many theoretical and practical areas. We introduce an innovative visual analytic approach to integrate microscopic and macroscopic perspectives of a rapidly growing scientific knowledge domain. Specifically, our approach focuses on statistically unexpected phrases extracted from unstructured text of titles and abstracts at the microscopic level in association with the magnitude and timeliness of their citation impact at the macroscopic level. The *H*-index, originally defined to measure individual scientists' productivity in terms of their citation profiles, is extended in two ways: 1) to papers and terms as a means of dividing these items into two groups so as to replace the less optimal threshold-based divisions, and 2) to take into account the timeliness of the impact of knowledge diffusion in terms of the timing of citations and publications so that attention is particularly drawn towards potentially significant and timely papers. The selected terms are connected to higher-level performance indicators, such as measures derived from the *H*-index, in the form of decision trees. A top-down traversal of such decision trees provides an intuitive walkthrough of concepts and phrases that may underline potentially significant but currently still latent scientific discoveries. Timeliness measures can also help to identify institutions that are at the forefront of a research field. We illustrate how widely accessible tools such as Google Earth can be utilized to disseminate such insights. The practical significance for digital libraries and fostering scientific discoveries is demonstrated through the astronomical literature related to the Sloan Digital Sky Survey (SDSS).

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries – *Collection*. H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Selection process*.

General Terms

Algorithms, Measurement, Experimentation.

Keywords

H-index, Sloan Digital Sky Survey, visualization, scientific discoveries, social networks

JCDL'07, June 17–22, 2007, Vancouver, British Columbia, Canada.
Copyright 2007 ACM 978-1-59593-644-8/07/0006...$5.00.

1. INTRODUCTION

The diffusion of scientific knowledge refers to the process in which scientific knowledge spreads through a scientific community or across disciplines. Identifying the significance of specific scientific concepts in such processes is essential to the understanding of the nature of scientific discoveries. One of the fundamental challenges is to integrate perspectives across different levels of granularity. In this article, we focus on associations between low-level features such as terms and phrases and high-level indicators of knowledge diffusion.

The focus of our study in this article is the scientific literature directly resulting from the Sloan Digital Sky Survey (SDSS[1]). The SDSS is the premier astronomical survey of our time. The SDSS Survey has provided numerous data to the astronomy community, including the approximately 10^6 brightest galaxies and 10^5 brightest quasars [40]. The early data release from the SDSS was in June 2001, including almost 14 million detected objects. It was followed by five official data releases annually since June 2003. The most recent one (DR5) contains 215 million unique objects. The SDSS data has become a real gold mine for astronomers. Important results of the SDSS include the discovery of a new class of stars within our Galaxy, the discovery of new galaxies orbiting the Milky Way, the discovery of the most distant quasars seen to date, at the edge of the observable universe, and the beginning of the measurement of dark energy and dark matter in the universe.

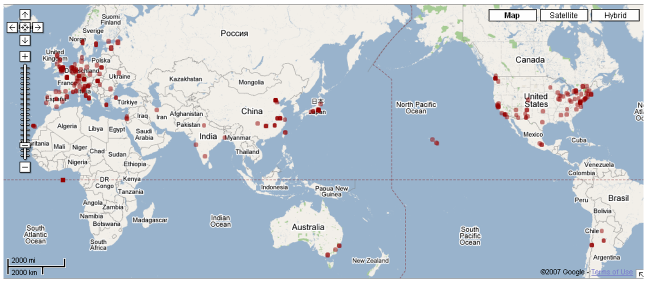

Figure 1. The SDSS data is used by astronomers all over the world. Each dot on the map marks the location of an author of an SDSS paper published between 2001 and 2006. The intensity of a marker indicates the frequency of publications in the corresponding geographic area.

The SDSS has led to an extremely fast-growing and high-impact body of literature. The SDSS literature currently includes nearly 1,400 papers with over 40,000 citations. The total citation number has doubled in the past 1.5 years. By citation impact, the SDSS

[1] http://www.sdss.org/

was the most important astronomical observatory in the world in 2003, 2004, and 2006 (it ranked second in 2005 to the NASA WMAP satellite). Michael Strauss, of Princeton University, recently posted a citation-ordered list[2] of all refereed papers to date with 'SDSS' or 'Sloan Survey' in their title or abstract. There are 89 papers with 89 or more citations; this is a generalized use of the *H*-index [19] on the SDSS literature.

2. CHALLENGES

Understanding even the most significant scientific discoveries in a fast-advancing field such as the SDSS is a challenging job. Investigating the massive volumes of observational data obtained by the SDSS survey is also a challenge that SDSS astronomers must deal with on a daily basis. Imagine the amount of effort it takes for SDSS astronomers to find their way through the interwoven data space and the knowledge space. The primary goal of our research in this area is to augment the ability of astronomers and information scientists to deal with the highly dynamic and transient body of scientific knowledge. We focus on establishing associative links between the massive volume of observational data and the most up-to-date scientific discoveries on relevant astronomical objects in scientific literature. Such links would enable astronomers to explore and investigate various emergent patterns across the data space and the knowledge space. Such links would also allow information scientists to study the interrelationship between the fast-growing data space and the evolving knowledge space and track the growth and spread of scientific knowledge at its forefront.

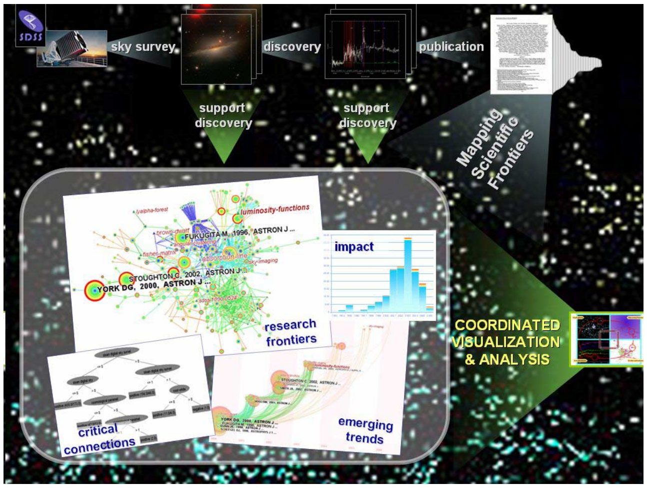

Figure 2. An overview of the SDSS literature analysis component of our approach.

We are developing an innovative approach that identifies and delineates statistically unexpected connections between phrases extracted from unstructured text and citation impact measures of a rapidly growing scientific knowledge space, especially including timeliness as well as magnitude of impact. The role of a phrase in the context of citation is depicted in an extended form of decision trees. A conceptual walkthrough of such decision trees is intuitive and extensible for further analysis. Timeliness of citations can help to identify institutions that are at the forefront of a research field. We are experimenting with Google Earth for conveying the diffusion of scientific knowledge.

Connecting text-level patterns and paper-level citations has not been done in a way that would give intuitive interpretations of the dynamics of a scientific field based on a growing body of text input and citation patterns. Making such connections is important for us to improve our understanding of science in the making. It is also important for the development of data mining and visual analytics algorithms. In addition to the astronomy and information science communities, broad impact is expected to reach other data-rich and fast-moving scientific fields. In this article, we introduce our approach to delineating the SDSS literature from an integrated microscopic and macroscopic perspective.

3. RELATED WORK

3.1 Studies of the Astronomical Literature

The astronomical literature in general has been extensively studied in the past. Abt addressed the issue of why some papers have long citation lifetimes [4]. He also studied trends in three leading American astronomical journals between 1910s and 1980s and found a continuous increase of annual publication rate since the WWII [3]. In 2000, Abt found that in the past decades "the number of published research papers worldwide show on abrupt change due to increased technical and scientific capabilities… The number of papers is a function only of the number of astronomers"[1]. He also noticed that the important papers did produce more citation than average papers [2]. Davoust and Schmadel [13] analyzed the worldwide astronomical publications from 1969 to 1987 and identified 14 "superproductive" astronomers, who published over 150 papers in 15 years.

Fernandez [16] compared single-author papers with multiple-author papers in two leading European astronomical journals between 1901 and 1996: *Astrophysical Journal* and the *Monthly Notice of the Royal Astronomical Society*. He found that the average number of authors per paper jumped from a little more than zero in the first half this century to about three in 1996. Fernandez's study confirmed that collaboratively authored papers are the mainstream of astronomical publication. On the other hand, multiple-author papers distribute differently among journals. We found a much higher average number of authors in the SDSS literature (9.5 between 2001 and 2006). More recently, Kurtz et al. [24] combined the NASA Astrophysics Data System (ADS) with astronomical journals, developing an easily accessible query-level digital database. ADS has comprehensive coverage of the SDSS literature.

3.2 Fostering Scientific Discoveries

Studies of scientific discoveries are distributed across a wide spectrum of disciplinary perspectives, ranging from studying scientific discoveries as a problem solving process [30, 32] to identifying the growth points in science guided by bibliographical statistics [25] to studying rapid theoretical changes through co-citation analysis [35].

Researchers in information science are increasingly challenged by the tension between the overwhelming volume of scientific literature and the lack of tools that can help them to uncover hidden structures across the boundaries of individual articles, to reveal how such structures evolve over time, and to understand what role is played by such structures in the advances of science [18].

Tabah [36] gives an insightful review of various issues concerning the dynamics of scientific literature, particularly the study of

[2] http://tinyurl.com/42jxy

growth, diffusion, and epidemics. Research in fields such as social network analysis [38], complex network analysis [6, 27, 29], and citation mapping and information visualization [8, 10, 33, 39] has produced a number of techniques that have the potential to tackle the structural complexity challenge.

Dunbar [14] studies the role of goal-setting strategies in making scientific discoveries in molecular biology. He concludes that goals are a powerful constraint on the cognitive processes underlying scientific reasoning and, more importantly, the types of goals influence the quality of reasoning. For example, he found that as soon as one notices evidence inconsistent with current hypotheses, a good strategy is to attempt to explain the cause of the discrepant findings.

Kostoff [22] described two methods that can be used to find potentially radical discoveries and innovations from “external” literatures, which can solve the problems defined in “internal” literatures. The two methods are the front-end component and the back-end component. The front-end method starts with a query to a discipline’s core literature, and then generalizes the query terms to search another discipline’s literature. The goal is to identify the potentially radical discoveries and innovations. The back-end method starts with the identification of radical discoveries and innovations from “external” sources, such as science and technology sponsoring organizations, journals, advisory panels, workshops, and review panels; it then finds links that bridge the external discoveries and innovation to the internal problems.

Daim et al. [12] utilized multiple methods, including patent analysis, bibliometric analysis, system dynamics, growth curves, and scenarios, to forecast the development of three technologies: fuel cells, food safety, and optical storage. Among those methods, bibliometric analysis was used to generate literature curves, which can demonstrate the development trail of a technology. Bibliometric analysis was also used to extract noun phrases, which formed the candidates of new technologies.

Ackermann [5] summarized the study of information epidemics in the scientific literature and identified six indicators:

1. Presence of one or a small group of seminal papers.
2. Rapid influx of numerous researchers who publish prolifically
3. Several distinct disciplines represented
4. Epidemic growth and decline of publication
5. Predominance of short communications published in rapid communication journals
6. Increase in multi-authorship of publications

Ackerman studied the Polywater and Cold Nuclear Fusion literature and found that indicators #1 through #4 are obvious in both cases, while indicators #5 and #6 are not.

Zitt et al. [41] introduced a hybrid method for extracting the bibliography of a specific scientific field, such as nanoscience. They first queried the ISI citation database to retrieve “seeds.” Next, they used an improved citation rule based on the “reference structure” function (RSF) to extend the “seeds” into a comprehensive collection in order to delineate a scientific subfield.

Temporal complexity refers to the temporal dynamics of scientific knowledge’s evolution and diffusion. How fast is fast for a new topic to receive citations? How is a new finding spreading within a scientific community? What is the most recent layer of the intellectual structure? What is the more recently formed sub-community? What was the turning point for the acceptance of a new theory? Research in knowledge discovery and data mining is particularly relevant, notably in the areas of concept drifts [21, 37], topic detection [26, 34], and change detection [20, 23].

An extensively used approach to understanding the dynamics of scientific knowledge and scholarly communication is to study the structure and dynamics of scientific literature. These studies of scientific literature can be further divided into two types based on whether they focus on text analysis or citation analysis [9]. Text analysis primarily focuses on the use of words and derives from similarities between different passages ranging from abstracts to entire full-text papers. Citation analysis, on the other hand, focuses on emergent patterns associated with references made by scientists in their publications. In contrast to word-frequency-based indexing mechanism commonly found in text analysis, citation indexing capitalizes on the potential intellectual value attributed to a referential link made by a scientist. A fundamental assumption is that such links are similar to a voting system by nature and they reflect a collective and contemporary view of many scientists on an intellectual association [33]. In practice, many researchers have pointed out that one should maintain caution when drawing conclusions, in particular if the specific citation context is not accessible.

In summary, our research aims to improve the understanding of the interrelationships between the advancement of science and the growth of its literature. The ongoing SDSS provides a good opportunity for the study of a rapidly growing body of scientific knowledge. A key objective of our work at this stage is to bridge the conceptual gap between local details in terms of phrases and terms used by scientists in their papers and the global structures of an evolving knowledge domain as perceived by the SDSS community. Specifically, our approach is to identify the role of information-rich terms in predicting the potential impact of underlying topics in the contemporary astronomy.

4. METHODS

Our method consists of several steps, including ranking papers by citation impact, categorizing papers by citation indices, selecting features from titles and abstracts of source records with reference to the citation impact of corresponding papers in which they appear, and generating and representing salient predictive relationships in decision trees.

4.1 Dataset

The SDSS literature dataset is provided by Thomson ISI, consisting of 1,350 bibliographic records, known as the source articles or records, written by 11,718 distinct authors. Source articles as a whole cited 25,946 references published between 1735 and 2007. The source articles themselves were collectively cited by 8,435 subsequently published articles between 1991 and 2006.

We removed two types of source records from the dataset: 1) records that are irrelevant to astronomy and 2) records that are relevant to the Sloan survey but purely focus on data releases and techniques of the survey. We removed the second type of records so that we can focus on scientific discoveries made with the SDSS data. A citation to these papers is required for papers using SDSS data; thus, their citations are not informative in this context. Seventeen data release-related records were removed for this reason.

The removal of the first type of records was due to the unexpected popularity of SDSS as an acronym. The source dataset was generated as a result of a search for SDSS. Since we expect a small number of pre-SDSS papers on the survey, we examined records dated before 1996 and manually removed records containing any of the following uses of SDSS as acronyms:

- SDS-Sedimentation
- Sodium dioctyl sulfosuccinate
- Self-Disclosure Situations Survey
- Spatial Decision Support Systems
- Strategic Decision Support Systems
- Superduplex Stainless Steels
- Superficial Dermatome Skin Samples

The remaining dataset contains 1,293 source records for subsequent analysis.

4.2 Measuring the Citation Impact

The number of citations to a published article is the most commonly used measure of the article's intellectual impact. It is easy to calculate and simple to understand. However, aggregating citations across a group of articles may not faithfully measure the impact of the group, especially when within-group differences are significant.

The *H*-index was originally proposed by Hirsch [19] as a measure of a scientist's scholarly productivity over his/her entire scientific career. It takes into account the number of publications and the number of citations of these publications and produces a simple metric. The *H*-index is defined as a number *h* for a scientist such that there are *h* papers published by the scientist with at least *h* citations, and the remaining publications by the same scientists have at most *h* citations. It effectively evaluates the output of a researcher from both impact and productivity. The *H*-index has been very popular in part because of its simplicity. However, the simplicity of the original *H*-index also limits its ability to track temporal variations of publications and citations more accurately.

Sidiropoulos et al. [31] generalized and normalized two variants of the *H*-index to reveal latent temporal facts in the citation networks; the contemporary *H*-index for "brilliant though young scientists who may have a small *h*" and the trend *H*-index for "trendsetters." Their research results inspire our experiments for exploring the potential of using expanded versions of the *H*-index as measures of citation impact.

For a scientist, one can consider his/her top *h* papers as the primary representatives of his/her work as far as the citation impact is concerned. Therefore, it would make sense if we can use the ideas and concepts expressed in these top *h* papers to characterize the nature of his/her citation impact. This notion of dividing papers into high- and low-impact groups can be easily extended to a collection or a digital library of papers written by different scientists. Dividing papers in this way is preferred to the use of arbitrarily chosen citation threshold values.

We use H_g to denote the generalized *H*-index. Given a collection of *N* papers ***C***, if there are H_g papers in the collection that have at least H_g citations and the remaining *N*- H_g papers have at most H_g citations, then H_g is the value of the generalized *H*-index for this collection of papers. The collection can be split into two subsets by H_g : $Split(\boldsymbol{C}, H_g) = \boldsymbol{S}_{\text{Low}} \cup \boldsymbol{S}_{\text{High}}$.

Following [31], we consider two scores to make adjustments to citation counts: S_t for a citation impact adjusted for timeliness, which gives heavier weights to citations made more recently than to citations made earlier, and S_c for a citation impact adjusted for publication age. S_t measures the impact in terms of the current citation trend, whereas S_c discounts citations accumulated over a long period of time. S_c and S_t scores are computed for each article *a* in the dataset according to [31]:

$$S_c(a) = \gamma * (Y_{now} - Y(a) + 1)^{-\delta} * |C(a)| \qquad (1)$$

$$S_t(a) = \gamma * \sum_{\forall c \in C(a)} (Y_{now} - Y(a) + 1)^{-\delta} \qquad (2)$$

where Y(*a*) is the year in which article *a* is published, Y_{now} is the current year, and C(*a*) is the set of articles that cite article *a*. γ is the coefficient and δ generally is set to 1.

As demonstrated by [31], the *H*-index can be extended further based on the two scores S_c and S_t as H_c and H_t. Similarly to the H_g extension we described above, we extend H_c and H_t to indices for a collection of papers from multiple authors. Again, the extended indices can be used to split such collections of papers. An H_t split will divide papers into new-born star papers and old-star papers, whereas an H_c split will separate papers into ones that have more time to collect citations and ones that are heavily cited within a short period of time. Thus, we have a number of ways to split a collection of papers into two groups based on an *H* split, an H_c split, and an H_t split in addition to splits based on average measures such as arithmetic mean and geometric mean. Multiple splits are an important component of our approach to contrast the role of concepts in different subgroups.

If the notion of average is defined for a group of entities, we often choose to separate them into the above-average group and the below-average group. Equations 3 and 4 define the arithmetic means and geometric means, where each term s_i is the S_t or S_c score of paper *i*. The arithmetic means and geometric means of the S_c and S_t scores are about 11 (See Table 2), except the geometric mean of S_t scores, which is 8.61. These numbers can be used to identify above-average papers and below-average papers in terms of their citation performance.

$$A(s_1, \ldots, s_n) = \frac{1}{n} \sum_{i=1}^{n} s_i \qquad (3)$$

$$G(s_1, \ldots, s_n) = \left(\prod_{i=1}^{n} s_i \right)^{\frac{1}{n}} \qquad (4)$$

4.3 Feature Selection

The feature selection step aims to select terms, i.e. phrases of multiple words, from the text fields of bibliographic records so that selected terms are content bearing and meaningful for our analysis and interpretation. First, we need to identify candidate terms. We utilize part-of-speech tagging to identify terms that contain up to four nouns with or without an adjective. Second, we want to select terms that would be most valuable in differentiating scientific discoveries reported in two subsets of the SDSS literature. For example, why is that one discovery is highly cited, but another one is not? By addressing questions like this, we could improve our understanding of the priorities and research agenda of SDSS research and attend to high-quality collections in a digital library more effectively.

We developed a feature selection method based on log-likelihood ratio [15] for differentiating conflicting opinions in customer reviews [11]. The extended H_g indices have made it possible for us to adapt the log-likelihood ratio method to select feature terms that are linked to citation impact measures, with additional adjustments for timeliness and freshness. Our selection method focuses on statistically unexpected connections. In essence, it tests statistical associations between terms and the category of the corresponding papers in which they appear.

Using log-likelihood ratios leads to a desirable advantage over threshold-based feature selection methods. That is, one can control the scope of a feature selection procedure by adjusting the statistical significance level (p-level) as a parameter. The lower the p-level, the more stringent the selection criteria, thus fewer associations are selected (See Table 1). In this study, the p-level is set to 0.01. Selected associations are used subsequently to build decision trees with terms as inner nodes and the categories of the split groups as leaf nodes.

Table 1. Statistical significance levels of log-likelihood ratios.

Log-likelihood ratio	p-level
15.13	0.0001
10.83	0.001
6.63	0.01
3.84	0.05

4.4 Decision Trees

Selected terms along with the citation status of hosting publications are used to generate decision trees so that the role of each term in the overall citation context of the SDSS can be represented in an easy-to-interpret form. The generation of decision trees is based on the concept of information gain to make a tree of classificatory decisions with respect to a previously chosen target classification. The information gain can be described as the effective decrease in entropy (usually measured in terms of bits) resulting from making a choice as to which attribute to use and at what level. For example, if one chooses a specified attribute like the length of a phase to discriminate among cases at a given point in its rule construction process, this choice will have some effect on how well the system can tell the classes apart. By considering which of the attributes is best for discriminating among cases at a particular node in the tree, we can generate a tree of decisions that allows us to navigate from the root of the tree to a leaf node by continually examining attributes. Decision trees classify every object within a dataset. Decision trees are often simplified by using pruning algorithms to reduce the size of the tree according to a user-defined level. In this study, we used a classic decision tree generating algorithm called C4.5 to generate decision trees because C4.5 is particularly suitable to meet our requirements in the following areas [28]:

- Avoiding overfitting the data
- Determining how deeply to grow a decision tree
- Handling continuous attributes
- Choosing an appropriate attribute selection measure

We also considered a variation of decision trees known as alternating decision trees (ADTrees). ADTrees are a generalization of decision trees by alternating layers of prediction nodes and splitter nodes [17]. As in decision trees, an instance is mapped to a path along the tree from the root to one of the leaves. On other hand, unlike decision trees, an ADTree maps each instance to a real valued prediction, which is the sum of the predictions of the path, rather than the label of a leaf. ADTrees extend decision stumps to represent classifications in real numbers rather than +1 or -1. For example, the classification of the paper containing the term *star formation* is sign (0.5-0.529)=sign(-0.029)= -1, which corresponds to the group of papers that have received above average citations. ADTrees tend to give smaller classification rules, which make them relatively easy to interpret.

5. RESULTS

The dataset contains 1,293 bibliographic records. The average title length is about 83 characters (82.96). The average abstract length is 1,389 characters (1,387.83).

5.1 H_g Indices and Splits

The 1,293-record dataset yielded an *H*-index of 65, including 3 papers with 65 citations each. The *H* split would put 67 papers in the highly cited group and the remaining 1,226 papers in the not highly cited group. Similarly, the H_c index for this dataset is 52, whereas the H_t index is 53.

We first split the dataset by the *H*-index, then we use the H_c and H_t splits, respectively. In addition, arithmetic or geometric mean is also used to divide the dataset. In this paper, however, we will focus on the geometric mean split and the *H* split.

Part-of-speech tagging identified a total of 22,665 terms, i.e. noun phrases. Our log-likelihood ratio method selected 290 terms. This is a significant reduction of the number of terms needed to be analyzed.

Table 2. Dividing the set of papers by arithmetic and geometric means.

	Sc	**Sc**	**St**	**St**
Total terms: 22,665	A(Sc)	G(Sc)	A(Sc)	G(Sc)
Pivotal value	11.70	11.06	11.46	8.61
#High	379	379	328	401
#Low	914	914	965	892

Figure 3 is a hybrid network visualized by CiteSpace, showing author assigned keywords (shown as concentric rings) and burst terms (shown as triangles) extracted from titles and abstracts. The concentric rings of a keyword depict the history of its use. For example, the term *spectroscopic target selection* has a red outmost ring, which indicates that it is a burst term. Rings in other colors correspond to individual time slices in the entire time interval (See the legend on top of the image). The color of its innermost ring indicates the earliest year this term first appears, in this case 2003. In contrast, the term early data release, located in the northeast quadrant of the image, has a thick ring corresponding to 2002, but two very thin rings corresponding to 2003 and 2004, respectively.

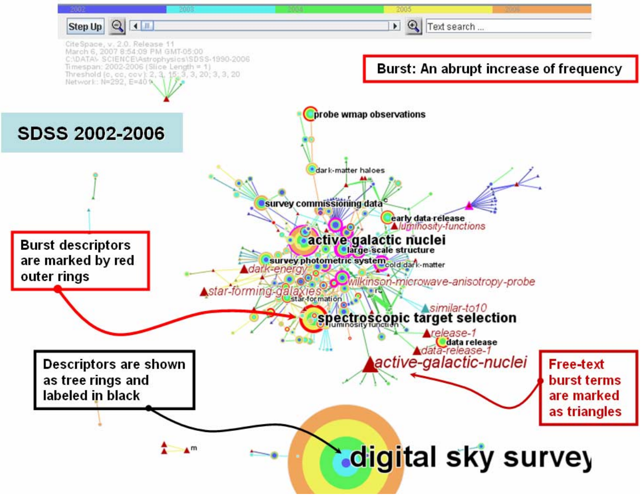
2002
2003
2004
2005
2006
Step Up
Text search ...
CiteSpace, v. 2.0. Release 11
March 6, 2007 8:54:09 PM GMT-05:00
C:\DATA\- SCIENCE\Astrophysics\SDSS-1990-2006
Timespan: 2002-2006 (Slice Length = 1)
Threshold (c, cc, ccv): 2, 3, 15; 3, 3, 20; 3, 3, 20
Network:: N=292, E=401
Burst: An abrupt increase of frequency
SDSS 2002-2006
probe wmap observations
dark-matter haloes
survey commissioning data
early data release
luminosity-functions
active galactic nuclei
large-scale structure
survey photometric system
cold dark-matter
dark-energy
wilkinson-microwave-anisotropy-probe
star-forming-galaxies
star formation
similar-to10
spectroscopic target selection
luminosity function
release-1
data release
data-release-1
active-galactic-nuclei
Burst descriptors are marked by red outer rings
Descriptors are shown as tree rings and labeled in black
Free-text burst terms are marked as triangles
digital sky survey

Figure 3. Prominent keywords assigned by authors and burst terms extracted from titles and abstracts (2002-2006).

5.2 Decision Trees

Figure 4 shows a decision tree of free-text terms selected by the log-likelihood ratio corresponding to the statistical significance level of p<0.01. The lower the p-level, the fewer the number of terms selected because of the more stringent selection criteria. The dataset was split by the geometric mean of the S_t scores. Figure 4 illustrates the overall structure of the tree rather than provide local details. Labels at this scale are too small to read. We include more readable examples in Figures 5, 6, and 7.

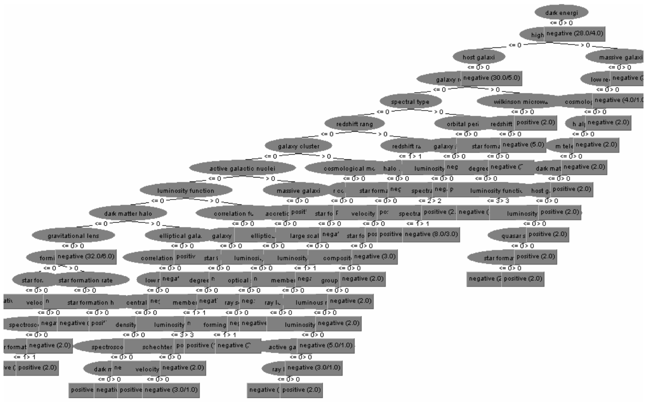

Figure 4. An overview of a decision tree generated based on 216 terms selected by log-likelihood ratio values (p<0.01) and a geometric mean split (74.44% of classification accuracy). The tree should be read from the root downwards (See also Figure 5).

Figure 5 shows a part of the tree in Figure 4 with larger-sized labels. The presence of the term *gravitational lens* is associated with the lower citation group, whereas terms on *star formation* tend to connect to the high and more timeliness citation group.

Figure 6 shows an ADTree derived from the same data split and the same feature set. The ADTree is much more compact than the earlier decision tree. For example, the left-most path shows that the term *dimensional power spectrum*, which is part of *three dimensional power spectrum*, is connected to the high-performance group because of the sign (0.409+-1.339<0).

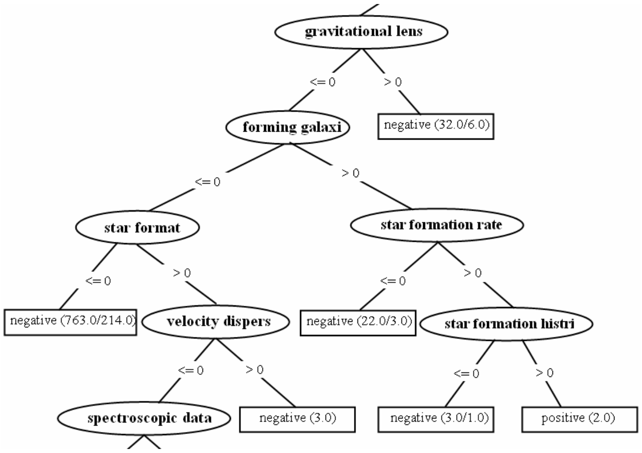

Figure 5. A part of the tree shown in Figure 4. The presence (>0) or absence (<=0) of a term is associated with a citation status group, i.e. highly and timely cited group.

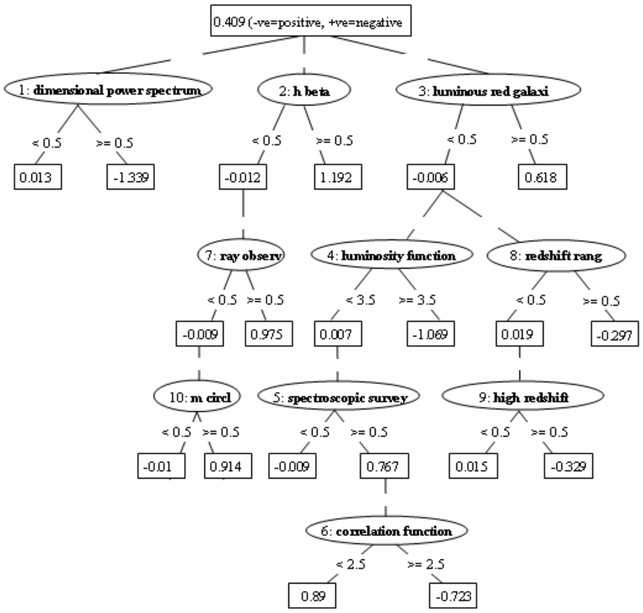

Figure 6. An ADTree derived from the data selected with the same selection criteria with 70.55% of accuracy.

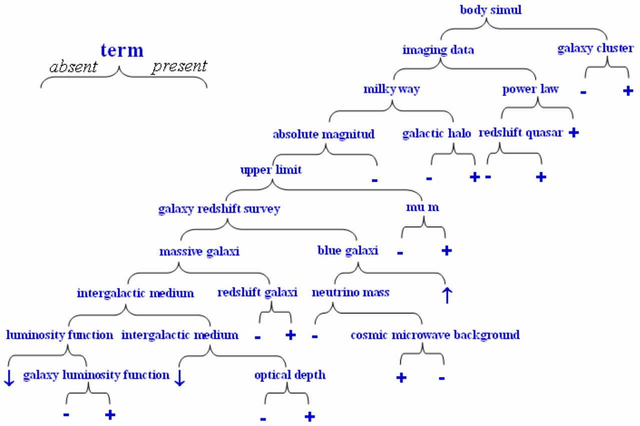

Figure 7. A decision tree of 95.82% classification accuracy derived from 721 terms and 1,267 records.

Figure 7 shows a decision tree derived from 721 terms and 1,267 bibliographic records. 95.82% of incidences are correctly classified by this decision tree (precision=0.769 and 0.962 for positive and negative groups, respectively; recall=0.299 and 0.995 for positive and negative groups, respectively).

5.3 Most Productive Organizations

The available citation impact indices make it possible to identify leading organization in terms of their accumulative productivity, timeliness of impact, and contemporariness of their research. Table 3 shows the most productive organizations in terms of the total number of papers they have produced in the entire dataset. Table 4 shows the top organizations in the high-performance group by the H-index split. Table 5 shows the top organizations in the high-performance group by the H_c split. Table 6 shows the top organizations in the high-performance group by the H_t split.

Table 3. Most productive organizations.

Organization	#Papers	City	Country
Princeton Univ Observ	71	Princeton	USA
Johns Hopkins Univ	39	Baltimore	USA
Univ Chicago	33	Chicago	USA
Univ Arizona	31	Tucson	USA
Univ Cambridge	31	Cambridge	England
Princeton Univ	29	Princeton	USA
Univ Washington	28	Seattle	USA
Ohio State Univ	27	Columbus	USA
Harvard Smithsonian Ctr Astrophys	25	Cambridge	USA
Penn State Univ	24	University Pk	USA
CALTECH	24	Pasadena	USA
Max Planck Inst Astrophys	23	Garching	Germany
NYU	22	New York	USA
Max Planck Inst Astron	21	Heidelberg	Germany
Univ Tokyo	20	Tokyo	Japan
Chinese Acad Sci	19	Beijing	Peoples R China
Fermilab Natl Accelerator Lab	17	Batavia	USA
Univ Pittsburgh	16	Pittsburgh	USA
Univ Durham	14	Durham	England
Univ Oxford	13	Oxford	England
Univ Calif Berkeley	13	Berkeley	USA
Univ Tokyo	13	Kashiwa	Japan
Los Alamos Natl Lab	11	Los Alamos	USA
Space Telescope Sci Inst	11	Baltimore	USA

Table 4. Top organizations in the high-performance group of the H-split (H=65).

#Papers	Organization	City, Country
9	Princeton Univ Observ	Princeton, USA
4	Fermilab Natl Accelerator Lab	Batavia, USA
4	Univ Durham	Durham, UK
3	Inst Adv Study	Princeton, USA
3	Johns Hopkins Univ	Baltimore, USA
3	Max Planck Inst Astron	Heidelberg, Germany
3	Max Planck Inst Astrophys	Garching, Germany
3	NYU	New York, USA
3	Penn State Univ	University Pk, USA
3	Univ Arizona	Tucson, USA
2	Univ Michigan	Ann Arbor, USA
2	Princeton Univ	Princeton, USA
2	Univ Penn	Philadelphia, USA
2	Ohio State Univ	Columbus, USA

Table 5. Top organizations in the high-performance group of the H_c split (H_c =52).

#Papers	Organization	City, Country
5	Princeton Univ Observ	Princeton, USA
4	Fermilab Natl Accelerator Lab	Batavia, USA
4	Max Planck Inst Astrophys	Garching, Germany
4	Univ Arizona	Tucson, USA
3	Johns Hopkins Univ	Baltimore, USA
3	NYU	New York, USA
2	Inst Adv Study	Princeton, USA
2	Max Planck Inst Astron	Heidelberg, Germany
2	Penn State Univ	University Pk, USA
2	Univ Durham	Durham
2	Univ Penn	Philadelphia, USA

Table 6. Top organizations by the H_t split (H_t=53).

Organization	#Papers	City	Country
Princeton Univ Observ	6	Princeton	USA
Fermilab Natl Accelerator Lab	4	Batavia	USA
Inst Adv Study	3	Princeton	USA
Max Planck Inst Astrophys	3	Garching	Germany
NYU	3	New York	USA
Univ Durham	3	Durham	England
Max Planck Inst Astron	2	Heidelberg	Germany
Penn State Univ	2	University Pk	USA
Univ Penn	2	Philadelphia	USA
Univ Michigan	2	Ann Arbor	USA
Ohio State Univ	2	Columbus	USA

Figure 8 is a screenshot of Google Earth in which we marked the locations of the first authors of SDSS papers. Using the H_c and H_t splits one can trace the changes of the most active organizations over time and space.

5.4 The Role of Timeliness Adjustments

The average year of citations in top 20 highly cited articles is 2001. The average year of publications in top 20 S_c ranked articles is also 2001, whereas the average year of publications in top 20 S_t ranked articles is 2002.

Figure 9 depicts the implications of the S_c and S_t scores on the adjusted citation curves. The original citation counts (citations) are shown as the highest line until the most recent two years. S_c, i.e. citations adjusted by the age of publication, is the lowest line until the most recent two years. The line in between is the S_t line, which tracks the raw citation line more closely than S_c.

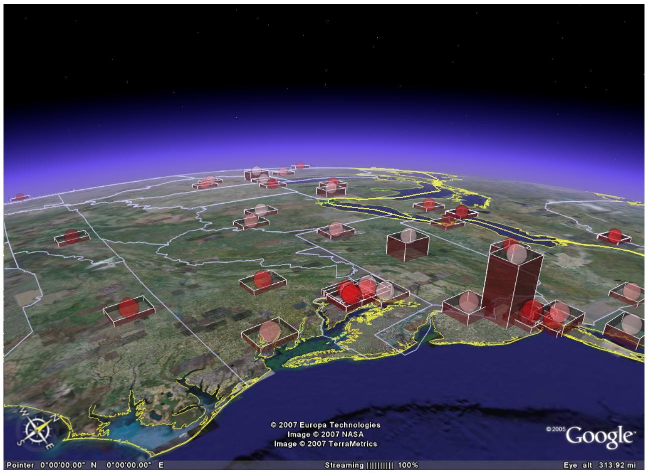
© 2007 Europa Technologies
Image © 2007 NASA
Image © 2007 TerraMetrics
©2005 Google
Pointer 0°00'00.00" N 0°00'00.00" E
Streaming ||||||||| 100%
Eye alt 313.92 mi

Figure 8. Timeliness can be used to select and track organizations. The view is facing west from the US east coast. The highest marker is located at Princeton.

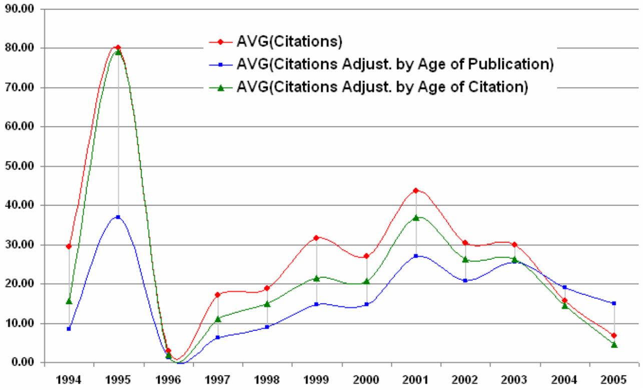

Figure 9. Citations adjusted by age of publication (S_c) and by age of citation (S_t).

Table 7. High impact papers by citations and S_t scores.

Year	Title	Cites	S_t
2004	Cosmological parameters from SDSS and WMAP	404	**367.00**
1995	THE FIRST SURVEY - FAINT IMAGES…	455	**301.64**
2003	Stellar population synthesis …	371	**263.47**
2001	Evidence for reionization at z similar to 6…	307	**255.07**
2001	The luminosity function of galaxies …	250	**196.73**
2003	A survey of z > 5.7 quasars …	195	**175.80**
2001	A survey of z > 5.8 quasars in the Sloan …	226	**174.87**
2002	Evolution of the ionizing background …	211	**170.00**
2001	Composite quasar spectra …	221	**168.21**
2004	The three-dimensional power spectrum …	224	**167.00**
2002	Observational constraints on growth of …	171	**155.93**
2002	Galaxy clustering in early Sloan …	157	**142.33**
2002	The 2dF Galaxy Redshift Survey: the b(J)-band …	160	**136.53**
2003	The galaxy luminosity function …	152	**126.87**
2001	High-redshift quasars found in Sloan …	151	**114.35**
2003	Stellar masses and star formation …	142	**113.27**
2001	Color separation of galaxy types in the Sloan …	127	**105.27**
2002	The 2dF Galaxy Redshift Survey …	111	**98.73**
2002	The ghost of Sagittarius and lumps …	128	**98.00**
2003	The dependence of star formation history …	121	**96.33**
2003	Galaxy star formation as a function of env…	113	**95.93**
2002	Toward spectral classification of L and T dwarfs	112	**88.00**
2003	The host galaxies of active galactic …	121	**87.33**
2000	The discovery of a luminous z=5.80 quasar …	123	**80.66**
2005	Cosmological parameter analysis including SDSS Ly alpha forest and galaxy bias…	134	**49.33**

Figure 10 shows the citation history of timeliness papers determined by their S_t scores. Darker lines correspond to the high-performance group, whereas lighter lines correspond to the other group. It shows that S_t scores tend to promote highly cited but more recently published papers.

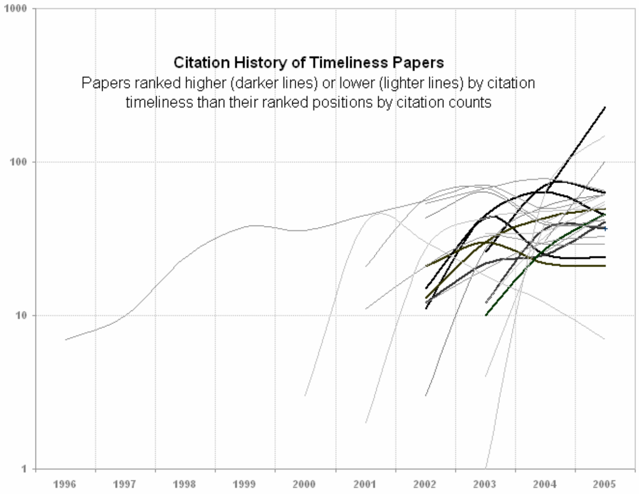

Figure 10. The citation history of timeliness papers shows recently published papers are moved up in the rankings.

6. DISCUSSIONS

Several issues require attention for further investigation. For example, term-level patterns give detailed insights into specific concepts. On the other hand, it would be valuable to provide additional layers of abstraction between the term level and the document level citation impacts. Terms in the following list are closely related to one another, but they also have subtle and yet distinct meanings.

observational data, observational discovery, observational estimate, observational evidence, observational measure, observational model, and observational result

A domain-specific ontological structure would be a logical step to capture complex relations between terms so that one can classify articles much more precisely than using term-level patterns alone.

An alternative approach to using existing ontological structures is to make use of the analysis of terminology variation patterns with which we experimented in differentiating conflicting opinions [11]. The strengths of terminology variations can be used as a grouping mechanism to construct higher-level aggregates. A significant advantage of the latter approach would be its domain-independent feature and its flexibility to be applied to different domains.

The timeliness-aware metrics are particularly valuable for detecting temporal patterns. They can be used to identify the very forefront of a scientific field, ranging from individual scientists and organizations.

The use of decision trees has generated some promising results. It is shown that the interpretation power has the potential to accommodate more complex and more sophisticated relations and dynamics.

7. CONCLUSIONS

We have shown that the *H*-index and some existing generalizations can be further extended from individual scientists to a collection of scientific papers or a digital library of a special knowledge domain. Furthermore, generalized metrics take into account temporal dimensions of the dynamics of a rapidly advancing scientific field. They provide additional options for dividing a collection of scientific papers into time- or performance-registered groups.

Terms extracted from hosting papers are a valuable source for arriving at insights into the interrelationship between term-level patterns at a microscopic level and citation impact patterns at a macroscopic level, such as a document-level, and even field-level.

Our experiment has generated promising results and inspirations for further investigation. The approach has the potential to contribute to the understanding of scientific discoveries at both theoretical and practical levels. Furthermore, this is an important step towards the establishment of tight conceptual links between scientific data and the relevant literature.

We suggest a few directions for future work, including unsupervised ontology construction to smooth the feature space, incremental classification of incoming new data and scholarly publications, self-directed optimization of existing decision trees based on new evidence, and full-text analysis that can model associative relations between hypotheses and evidence and between facts and opinions.

Digital libraries should provide scientists not only with well-organized and accessible scientific literature but also with intellectual pathways that can lead to scientific discoveries and knowledge creation, trailblazing and transforming the knowledge space as envisaged by Vannevar Bush in his Memex [7] .

ACKNOWLEDGMENTS

This work is supported by the National Science Foundation under Grant No. 0612129. Thomson ISI provides the bibliographic data for the analysis.

REFERENCES

[1] Abt, H.A. Astronomical publication in the near future. *Publications of the Astronomical Society of the Pacific*, *112*, (2000), 1417-1420.

[2] Abt, H.A. Do important papers produce high citation counts. *Scientometrics*, *48*, (2000), 65-70.

[3] Abt, H.A. Some trends in American astronomical publications. *Publications of the Astronomical Society of the Pacific*, *553*, (1981), 269-272.

[4] Abt, H.A. Why some papers have long citation lifetimes. *Nature*, *395*, (1998), 756-757.

[5] Ackermann, E. Indicators of failed information epidemics in the scientific journal literature: A publication analysis of Polywater and Cold Nuclear Fusion. *Scientometrics*, *66*, 3 (2006), 15.

[6] Barabási, A.L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A. and Vicsek, T. Evolution of the social network of scientific collaborations. *Physica A*, *311*, (2002), 590-614.

[7] Bush, V. As we may think. *The Atlantic Monthly*, *176*, 1 (1945), 101-108.

[8] Chen, C. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, *57*, 3 (2006), 359-377.

[9] Chen, C. *Mapping Scientific Frontiers: The Quest for Knowledge Visualization*. Springer, London, 2003.

[10] Chen, C. Searching for intellectual turning points: Progressive Knowledge Domain Visualization. *Proc. Natl. Acad. Sci. USA*, *101*, Suppl. (2004), 5303-5310.

[11] Chen, C., Ibekwe-SanJuan, F., SanJuan, E. and Weaver, C., Visual Analysis of Conflicting Opinions. in *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST)*, (Baltimore, MA, 2006), 2006, 59-66.

[12] Daim, T.U., Rueda, G., Martin, H. and Gerdsri, P. Forecasting emerging technologies: Use of bibliometrics and patent analysis. *Technological Forecasting and Social Change*, *73*, (2006), 31.

[13] Davoust, E. and Schmadel, L.D. A study of the publishing activity of astronomers since 1969. *Scientometrics*, *22*, (1991), 9-39.

[14] Dunbar, K. Concept discovery in a scientific domain. *Cognitive Science*, *17*, (1993), 397-434.

[15] Dunning, T. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, *19*, 1 (1993), 61-74.

[16] Fernandez, J.A. The transition from an individual science to a collective one: The case of astronomer. *Scientometrics*, *42*, (1998), 61-74.

[17] Freund, Y. and Mason, L. The alternating decision tree learning algorithm *Proceeding of the Sixteenth International Conference on Machine Learning*, Bled, Slovenia, 1999.

[18] Hargens, L.L. Using the Literature: Reference Networks, Reference Contexts, and the Social Structure of Scholarship. *American Sociological Review*, *65*, 6 (2000), 846-865.

[19] Hirsch, J.E. An index to quantify an individual's scientific research output. *PNAS*, *102*, (2005), 16569-16572.

[20] Kleinberg, J., Bursty and hierarchical structure in streams. in *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (Edmonton, Alberta, Canada, 2002), ACM Press, 2002, 91-101.

[21] Klinkenberg, R. and Renz, I., Adaptive information filtering: learning in the presence of concept drifts. in *Learning for Text Categorization*, (Menlo Park, CA, 1998), AAAI Press, 1998, 33-40.

[22] Kostoff, R.N. Systematic acceleration of radical discovery and innovation in science and technology. *Technological Forecasting and Social Change*, *73*, (2006), 13.

[23] Kumar, R., Novak, J., Raghavan, P. and Tomkins, A., On the Bursty Evolution of Blogspace. in *WWW2003*, (Budapest, Hungary, 2003), 2003, 477.

[24] Kurtz, M.J., Eichhorn, G., Accomazzi, A., Grant, C.S., Demleitner, M. and Murray, S.S. Worldwide use and impact of the NASA Astrophysics Data System digital library. *Journal of the American Society for Information Science and Technology*, *56*, (2005), 36-45.

[25] Meadows, A.J. and O'Connor, J.G. Bibliographical statistics as a guide to growth points in science. *Science Studies*, *1*, 1 (1971), 95-99.

[26] Morinaga, S. and Yamanishi, K., Tracking dynamics of topic trends using a finite mixture model. in *KDD'04*, (Seattle, Washington, 2004), ACM, 2004, 811-816.

[27] Newman, M., The structure of scientific collaboration networks. in *Natl. Acad. Sci*, (USA, 2001b), 2001b, 404-409.

[28] Quinlan, R. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA., 1993.

[29] Radicchi, F., Castellano, C., Cecconi, F., Loreto, V. and Parisi, D. Defining and identifying communities in networks *arXiv: cond- mat/ 0309488 v1*, 2003.

[30] Root-Bernstein, R.S. *Discovering: Inventing and solving problems at the frontiers of scientific knowledge*. Harvard University Press, Cambridge, 1989.

[31] Sidiropoulos, A., Katsaros, D. and Manolopoulos, Y. Generalized h-index for disclosing latent facts in citation networks. *arXiv:cs.DL/0607066* (2006).

[32] Simon, H.A., Langley, P.W. and Bradshaw, G.L. Scientific discovery as problem-solving. *Synthese 47*, (1981), 1–27.

[33] Small, H. Visualizing science by citation mapping. *Journal of the American Society for Information Science*, *50*, 9 (1999), 799-813.

[34] Steyvers, M., Smyth, P., Rosen-Zvi, M. and Griffiths, T., Probabilistic author-topic models for information discovery. in *KDD'04*, (Seattle, Washington, 2004), ACM, 2004, 306-315.

[35] Sullivan, D., Koester, D., White, D.H. and Kern, R. Understanding Rapid Theoretical Change in Particle Physics: A Month-By-Month Co-Citation Analysis. *Scientometrics*, *2*, 4 (1980), 309-319.

[36] Tabah, A.N. Literature dynamics: studies on growth, diffusion, and epidemics. *Annual Review of Information Science and Technology*, *34*, (1999), 249-286.

[37] Tsymbal, A., Pechenizkiy, M., Cunningham, P. and Puuronen, S. Dynamic integration of classifiers for tracking concept drift in antibiotic resistance data *Technical Report TCD-CS2005-26*, Department of Computer Science, Trinity College, Dublin, Ireland, 2005.

[38] Wasserman, S. and Faust, K. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[39] White, H.D. and McCain, K.W. Visualizing a discipline: An author co-citation analysis of information science, 1972-1995. *Journal of the American Society for Information Science*, *49*, 4 (1998), 327-356.

[40] York, D.G., Adelman, J., Anderson, J.E., Anderson, S.F., Annis, J., Bahcall, N.A. and al., e. The Sloan Digital Sky Survey: Technical summary. *Astronomical Journal*, *120*, (2000), 1579-1587.

[41] Zitt, M. and Bassecoulard, E. Delineating complex scientific fields by an hybrid lexical-citation method: An application to nanosciences. *Information Processing & Management*, *42*, (2006), 18.