Identifying Thematic Variations in SDSS Research

Chaomei Chen[1], Fidelia Ibekwe-SanJuan[1,3], Eric SanJuan[2], Michael S. Vogeley[4]

[1]iSchool, Drexel University – 3 141 Chestnut Street – Philadephia, PA 19 104 – USA
{chaomei.chen@cis.drexel.edu}

[2] LIA, Université d'Avignon – 339, Chemin des Meinajaries
84 911 Avignon CEDEX 09, France {eric.sanjuan@univ-avignon.fr}

[3] ELICO – Université de Lyon3, 69 008 Lyon. {ibekwe@univ-lyon3.fr}

[4]Department of Physics, Drexel University – 3 141 Chestnut Street – Philadephia,
PA 19 104 – USA. {vogeley@drexel.edu}

Abstract

The Sloan Digital Sky Survey (SDSS) is the largest ongoing sky survey. It regularly makes data releases to the astronomical community. From a macroscopic point of view, a profound question is: what is the role of SDSS data releases in the evolution of the relevant scientific fields? In this paper, we introduce an integrated approach by combining statistical, information-theoretical, and symbolic methods for text data analysis and show how this combined approach can distinguish thematic variations associated with the different data releases.

Keywords: text mining, topic detection, information visualization, thematic evolution.

Résumé

Le projet scientifique « Sloan Digital Sky Survey » (SDSS) est actuellement le plus grand projet d'observation des objets célestes pour la communauté de l'astronomie. D'un point de vue macroscopique, une question fondamentale se pose : celle de savoir comment les données astronomiques observées et rendues publiques par le projet SDSS influent sur l'évolution de la recherche scientifique en astronomie. Comment les chercheurs se servent-ils des données observées pour orienter leurs recherches ? Nous proposons une approche intégrée de la fouille de textes qui combinent des méthodes statistique et symbolique pour démontrer comment cette approche combinée est à même de répondre aux questions de l'évolution de la recherche en astronomie.

Mots-clés: fouille de textes, détection de thèmes, visualisation de l'information, évolution des thèmes.

1. Introduction

Discovering and tracking important themes in a vast volume of data is a fundamental challenge for several text mining tasks (Berry 2004), including science and technology watch (Schiffrin & Börner 2004; Ibekwe-SanJuan & SanJuan 2004) and mapping scientific frontiers (Chen, 2006). As the premier astronomical survey of our time, the Sloan Digital Sky Survey (SDSS) has generated a rich and rapidly growing body of literature on its discoveries. The SDSS survey aims to provide detailed optical images covering more than a quarter of the sky, and a 3-dimensional map of about a million galaxies and quasars.

In this study, we apply two different text mining systems to identify thematic trends in papers citing SDSS data releases, namely CiteSpace II (Chen 2003, 2006) and TermWatch (SanJuan & Ibekwe-SanJuan 2004, 2006). Both systems are designed for clustering information units,

such as authors, author assigned keywords, noun phrases, at several levels of granularity. The systems also include a visual analytical component so that the user may explore and assimilate complex relations between information units. In this study, CiteSpace is used to extract, select and visualize associations between pairs of document features. This analysis of association graphs is then reinforced by generalizing binary associations with frequent item sets for each individual SDSS data release on terms extracted by TermWatch. Finally, we use atom graph decomposition implemented in TermWatch to select and visualize frequent item sets.

Our corpus consists of 1,293 bibliographic records of SDSS-related publications retrieved from the "ISI Web of knowledge" (Thomson scientific) database. These records contain the usual bibliographic fields such as title, author, abstract, keywords, affiliation and date. Since our goal is to distinguish the impact of individual SDSS data release, in addition to the analysis of the entire corpus. The entire dataset was divided into several subsets corresponding to individual SDSS data release. Each data release is described by a major technical paper. In this study, we analyze the first 6 data releases. The last one DR6, released in 2007 was omitted.

The rest of the paper is structured as follows: sections §2 and §3 present an overview of the CiteSpace and TermWatch systems respectively with particular emphasis on the features used to mine the SDSS corpus. Section §4 analyzes results and section §5 draws some conclusions from the experiment.

2. Building association graphs with CiteSpace

CiteSpace is designed for visualizing and analyzing emerging trends and macroscopic changes of a scientific domain through its scientific literature (Chen 2004; Chen 2006). The overall design is depicted in the following diagram (Figure 1). Typically, CiteSpace takes a set of bibliographic records as its input and generates interactive visualizations that help users to explore and identify macroscopic patterns. For technical details, see (Chen, 2004 and Chen, 2006). CiteSpace first divides the incoming records into a series of time slices. Data in each time slice is processed and a network is generated. A network consists of two types of entities, namely nodes and links. CiteSpace supports author, article, journal, institution, and country nodes. It also supports phrases extracted from titles and abstracts of articles. We will particularly focus on the role of phrases in this paper. CiteSpace supports link types that can be derived from given node types. For example, links between a phrase and an article are defined as referential links, whereas collaborative links are derived from co-authorship. Once networks in individual time slices are generated, these networks will be merged to form a global network. A number of filters are available to the global network so that one can render the network with more specific focus. For example, betweenness centrality measures (Freeman, 1973) of nodes are used to identify pivotal points of intellectual change. In this paper, we will focus on a few statistical filters which are relevant to the task at hand.

The features of CiteSpace used in this study include i) burst detection (Kleinberg 1999), ii) feature selection based on log-likelihood ratio tests of statistical association, and iii) information-theoretic indices (Chen 2008). All three features are applied to multi-word phrases extracted from the SDSS literature corpus. They will be detailed in the results section (§4).

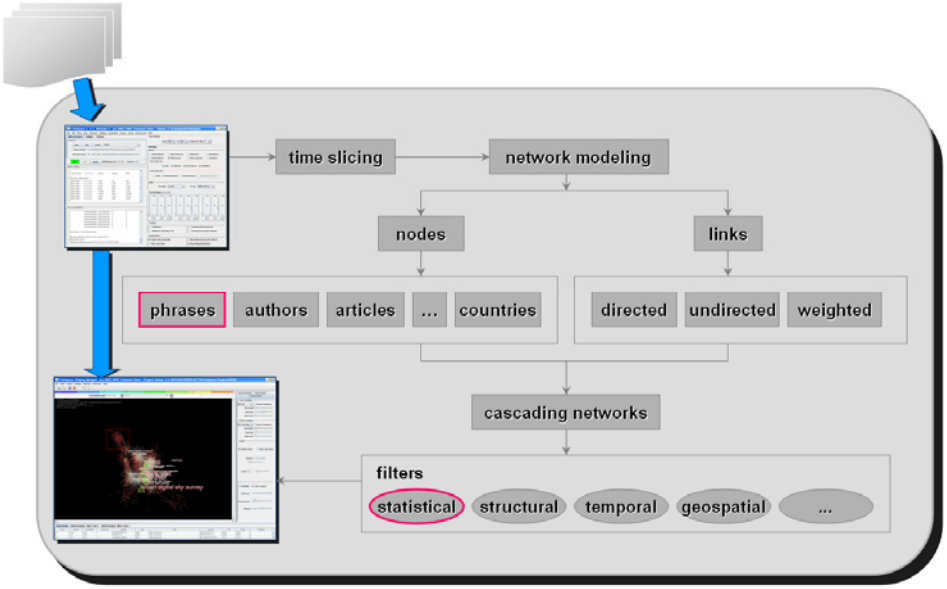

Figure 2. An information flow diagram of CiteSpace.

3. Mining frequent item sets and atom graphs with TermWatch

First, we give a general overview of TermWatch before describing in more details the specific features used in this experiment.

TermWatch (SanJuan & Ibekwe-SanJuan 2004, 2006) maps research topics at the term level. It relies on surface linguistic and terminological information to extract terms and construct a graph of terminological relations between them. We recently added some new text mining features such as frequent item set mining based on "arules R package" and atom graph decomposition. These recent additions to the system are the main functionalities used in the current experiment. Frequent item set generalize the idea of associations. They can point *n*-ary relations between terms whereas associations induce a binary relation. We first use CiteSpace to obtain a macroscopic view based on binary relations on concepts, and then use TermWatch's term extraction and frequent item sets mining to reveal specific data release themes. Figure 2 below gives a pipeline view of the different levels of text analysis in TermWatch.

3.1. Term extraction and feature selection

Usually, in TermWatch's term extraction procedure, multi-word terms (without limit of length) are extracted based on their morpho-syntactic properties. All extracted terms are used in further processing without any regard for occurrence. However, in this exsperiment, the focus was not on rare items, hence we had to devise a feature selection index which would eliminate rare terms. First, we carried out term extraction using morpho-syntactic patterns and eliminated all terms that only occured once. We then computed the geometric mean of the inertia induced by two tf.idf functions. One is based on the whole term occurrence, the other is based on the occurrence of term's words using MYSQL match function and document length normalization. The topmost terms must be frequent, so that they are specific to a subset of documents without being too frequent or evenly distributed across the corpus.

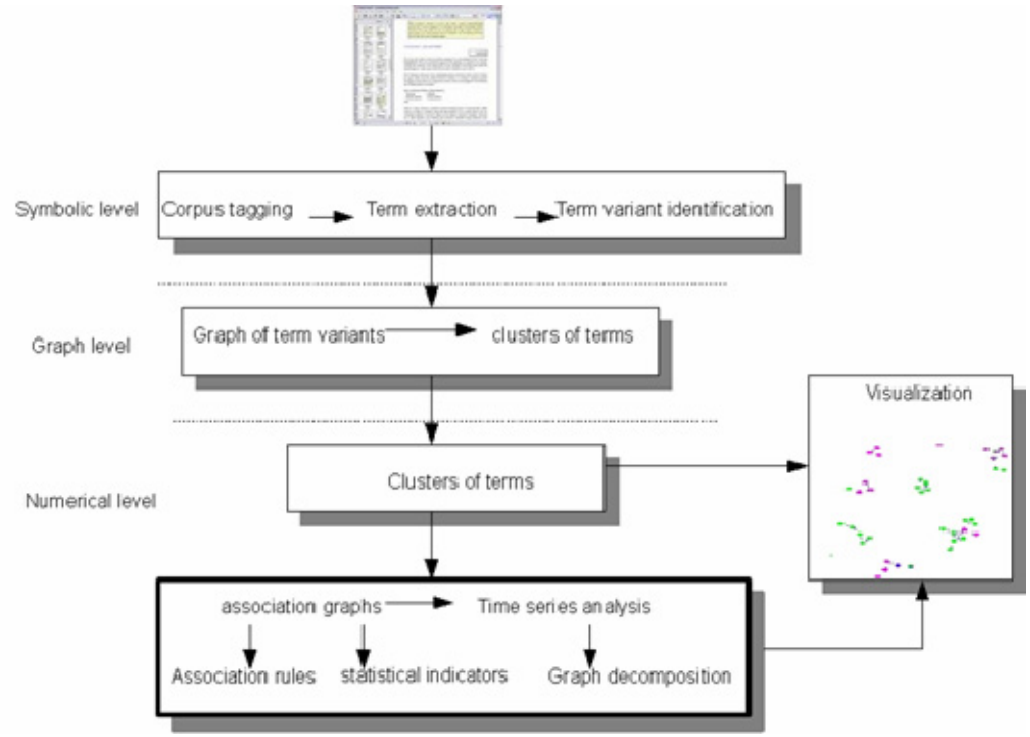

Figure 3. A pipeline view of text analysis features in TermWatch.

Candidate Terms are thus ranked according to a score G(t) that is computed in the following way:

$$G(t) = \frac{F(t) \times V(t)}{max\{F(t) : t \in T\} \times max\{V(t) : t \in T\}}$$

where:

$$F(t) = \sqrt{\sum f_{t,d}^2 \times \ln(|\{d : t \in d\}|)}$$

and

F(t) favors named entity meanwhile V(t) criteria favors more general terms like technical approaches or methods.

$$V(t) = \sqrt{\sum (\text{match } d \text{ against } t)^2}$$

Table 1 hereafter gives the first 20 terms ranked by G-Mean.

1 G_Mean	tf.idf	Match_SQL_	Freq	Docs	term
0,55	0,95	0,57	78	43	white dwarf
0,38	0,59	0,65	49	39	dark matter halo
0,36	1	0,36	89	64	luminosity function
0,32	0,72	0,45	57	38	velocity dispersion
0,32	0,81	0,39	74	58	star formation
0,3	0,63	0,47	59	53	active galactic nucleus
0,28	0,77	0,36	62	40	early-type galaxy
0,28	0,59	0,47	50	42	large-scale structure
0,27	0,43	0,63	37	35	cosmic microwave background
0,26	0,52	0,5	39	28	star formation history
0,26	0,61	0,42	45	29	power spectrum
0,26	0,62	0,41	56	49	emission line
0,25	0,74	0,34	59	38	stellar mass
0,25	0,39	0,63	27	18	black hole mass
0,24	0,7	0,34	60	46	spectral type
0,24	0,5	0,48	42	36	dark matter
0,24	0,43	0,55	30	19	black hole
0,23	0,47	0,49	37	30	gravitational lensing
0,21	0,32	0,65	23	18	supermassive black hole
0,21	0,62	0,33	48	33	photometric redshifts

Table 1. First 20 terms ranked by G_mean index.

3.2. Frequent item sets

We now work on the relation between documents and their terms. In order to clarify the connections between frequent item sets mining and association graphs, we shall see any set D of documents as a hyper-graph H. Each document is represented by a hyper-vertex. Each of the elements of text data, for instance terms, is represented by the subset of documents sharing this particular attribute. These subsets form the hyper-edges of the hyper-graph.

Formal Concepts correspond to hyper-graph minimal transversals. Indeed, for any subset S of hyper-vertices, let us denote by $T(S)$ the set of hyper-edges h such that $h \cap S \neq \emptyset$. S is said to be a minimal transversal if for any s in S, $T(S - s) \neq T(S)$. Now, we consider the relation R between the set D of documents and the set K of terms defined by $(d, t) \in R$ if and only if document d contents term t $(t \in d)$.

Given a set of keywords V, let us denote by Ext(V) the set of documents indexed by all terms in V. Ext(V) is called the formal extension of V. Conversely, given a set of documents U, let us denote by $Int(U)$ the set of terms indexing all documents in U. $Int(U)$ is called the formal intension of U. Thus a formal concept is a pair (U, V) such that Ext$(V) = U$ and Int$(U) = V$. It is straightforward to check that if S is a minimal transversal, then $(S, T(S))$ is a formal concept. Formal concepts are partially ordered by inclusion on their extensions. The result is a lattice. A subset of keywords or terms is said to be closed if it is the intension of a formal concept or equivalently, a minimal transversal of the hyper-graph.

Depending on the type of information that the user wants to analyze, knowledge maps can be derived form subparts of H as intersection graphs, that we shall also call association graphs and denote them by $G_k(V,E)$ where V is a set of vertices (nodes), E is a set of non-directed edges (links) and k is an integer. The vertices in V represent the selected hyper-edges while an edge is drawn among two vertices (w_1,w_2) whenever they have at least k elements in their intersection.

For small values of k, $G_k(V,E)$ is expected to be a Small World Graph (SWG). A graph is said to be SWG when it simultaneously shows both low diameter and high clustering measure, (i.e., high density of edges in the neighborhood of each vertex). According to (Ferrer & Sole

2001), the path length *L(p)* and the clustering coefficient *C(p)* are the two structural measurements characterizing SWGs. The usual approach for visualizing a SWG consists in computing a decomposition into highly connected components and offering the user an abstract view of the network to start with (Auber et al. 2003).

We adopt a similar approach except that we compute overlapping atoms (Berry et al., 2004) instead of disjoint connected components. The atoms of a graph can be defined based on the concept of (a, b)-clique separators. These are complete subgraphs (all vertices are related) such that there exists two vertices *a, b* not in the separator and such that any path from *a* to *b* necessarily contains at least one element in the separator.

We shall say that a graph is inseparable if there is no subgraph that is a complete separator. We shall call atom of a graph any connected maximal subgraph that is inseparable. By definition an atom A of G_k contains at least one complete separator S of G_k. However S is not a separator of A. Atoms overlap if they contain the same separator of Gk. The decomposition of G_k in atoms is unique and it can be decomposed in $O(|V| \times |E|)$.

In our experiments, we have observed that graphs of the form $G_k(V,E)$ for k between 1 and 3 have a central atom with long cycles that involves almost 50% of the vertices and numerous peripheral atoms of small size that are almost chordal (circles have less than three elements).

3.3. Atom graphs

To visualize atoms and their interactions on a map, we define a valued graph exclusively based on the structure of $G_k(V,E)$. Each atom A is labeled by the vertex w_1 having the highest degree defined as the number of edges linking w_1 to another vertex w_2 in A. Atoms having the same label are merged together.

The valued graph of atoms that we shall denote by $G_k(At) = (V_{At}, E_{At}, a_{At})$ is defined as follows.

The vertex of $G_k(At)$ are pairs of the form (k, l) where k is a vertex of G_k and l is the label of an atom containing k. An edge $e = (w_1, w_2)$ is defined between two vertices $w_1 = (k_1, l_1)$ and $w_2 = (k_2, l_2)$ if one of the following happens:

- $l_1 = l_2$ and (k_1, k_2) is an edge of G_k. In this case the value s_{At} of the edge e is set to 1.
- $k_1 = k_2$ and there exists a clique separator S in G_k that separates the atom l_1 from the atom l_2. In this case $a_{At}(w1,w2)$ is set to the ratio between the number of elements in S and the total number of elements in atoms l_1 and l_2.

The first case corresponds to edges in atoms. To ensure that the related vertices will not be separated by any clustering procedure, we set the value of such edges to 1, the maximum. The second case deals with edges relating copies of G_k vertices in different atoms. This valued graph can be displayed as described here below.

Finally, using the interactive interface AiSee (http://www.aisee.com) and its optimized bi-scale force directed layout, we obtain a two level access to the network of keywords. The idea is to enhance the visualization of the graphs obtained by TermWatch by identifying a core network of topics which form an inseparable sub-graph, and distinguish it from other satellite or peripheral research topics.

AiSee needs as input a file in Graph Description Language (GDL). Our GDL generator uses edge width to visualize the strength of the link. Clusters are then represented by ovals whose size depends on the number of clustered vertices. Finally, clusters can be unfolded in a wrapped form to visualize the transitions to other clusters.

4. Results

Several maps were obtained from the SDSS corpus both at the macroscopic high level view on the whole corpus (author co-citation networks) and the microscopic level. For reasons of space, we will focus here on the temporal evolution of topics across the data releases (DR). Also, this aspect is an important concern of the SDSS literature survey project.

4.1. Association network

CiteSpace computed the strength of associations between extracted terms using the log-likelihood ratio tests. The primary motivation of using log-likelihood ratio tests is due to its strengths in modeling the behavior of low-frequency text units. Associations between terms are selected if they are statistically significant at the level of p=0.01. Figure 3 shows a network visualization of terms and their associations selected in this way.

The lighter shades denote terms found in the earlier period of the corpus. Darker shades denote terms found in the later period of the corpus. For example, terms such as "*star formation histories, x-ray group, and group membe*r" form a light-colored small cluster. In contrast, "*x-ray source and upper scorpiu*" are connected in the later period of the corpus.

4.2. Burst terms

According to Kleinberg (1999), burst terms are text units that appear more often than expected in a particular time frame. Hence their detection can point to a surge in interest in a topic at that time. CiteSapce was used to detect burst terms across all DR datasets. The corresponding image is shown in Figure 4. We annotated the date of a particular data release. The early data release (EDR) took place in June 2001 and seems to be focused on "*imaging data*", a burst term prominent in that period. Later studies started referring to the EDR, hence the surge of the term "*early data release*" detected in the following year 2002. The first data release (DR1) was released in early 2003 with the surge of terms such as "*redshift range, dark matter*". References to DR1 by other researchers became significant in 2004, hence the surge of the term "*SDSS-DR1*" in this year along with other co-occurring terms such as "*spatial distribution*" and "*color distribution*". The second data release (DR2) became available in 2004 and the reference to "*DR2*" became statistically significant in 2005 with the term "*stellar metallicity"*. The impact of DR3 in terms of associated burst terms appears to relate to *quasar spectra*. DR4 is marked by the following burst terms: "*correlation function, velocity dispersion, galaxy-evolution explorer*".The reference to DR4 became significant in DR5 alongside other burst terms such as *lambda-lambda, quasar spectra*.

Figure 3. A network of keywords based on the strengths of association measured by the log-likelihood ratios. Keywords with lighter colors appeared in the early years. The entire SDSS dataset is used.

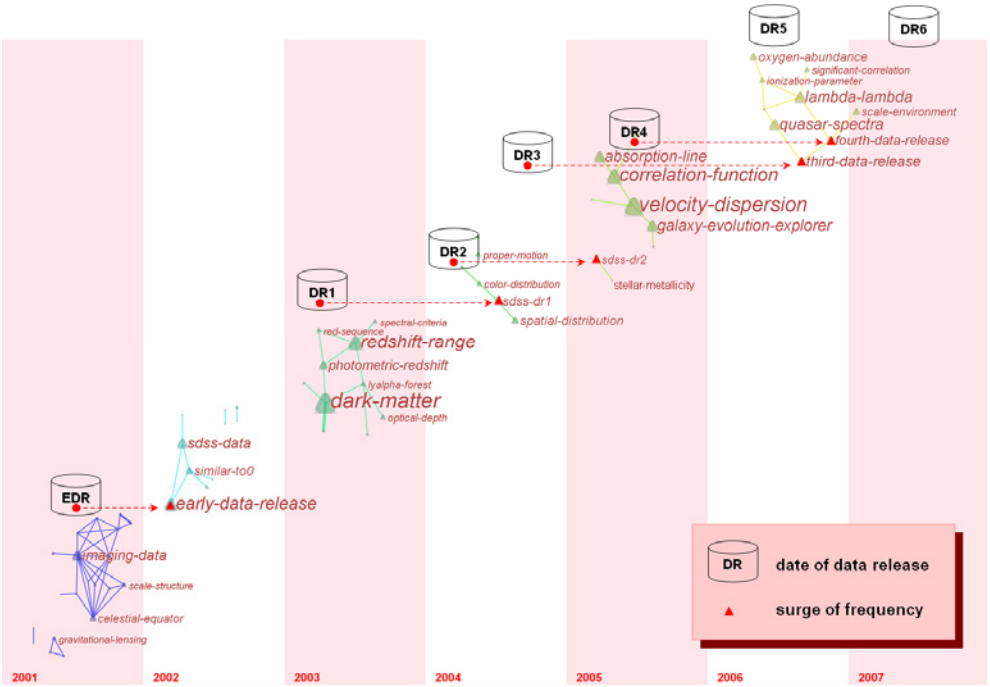

Figure 4. A timezone view generated by CiteSpace showing burst terms from the SDSS literature dataset (1993-Oct 2007). The image is annotated with the date of each SDSS data release.

4.3. Mining frequent item sets across data releases

In order to get an in depth view of topic evolution along the different data releases, we computed frequent item sets and atom graphs on the basis of terms extracted by TermWatch on the whole corpus but distributed among the papers referring to the different DRs. For reasons of space, we limit our results to rules and atom graphs obtained on records citing DR1 and DR2. The graphs below show the association rules mined from terms in DR1.

4.3.1. Association rules and atom graphs based on Data Release 1

For DR1, the 10 top-most frequent association rules are discussed in the following. We represent them under the form of association rules since in these frequent item sets, the appearance of a third or fifth element can be inferred from the others. Thus each of the following rules corresponds to a frequent item set made of all terms in the rule (left right part).

{equatorial strip, volume-limited subsamples} => {equatorial plane}
{equatorial plane, volume-limited subsamples} => {equatorial strip}
{equatorial strip, volume-limited subsamples} => {galaxy distribution}
{model atmosphere, white dwarf} => {first datum release}
{equatorial strip, volume-limited subsamples} => {galaxy distribution }
{volume-limited subsamples} => {equatorial plane}
{large number, proper-motion catalog} => {USNO-B}

Figure 5. Top-10 association rules related to DR1.

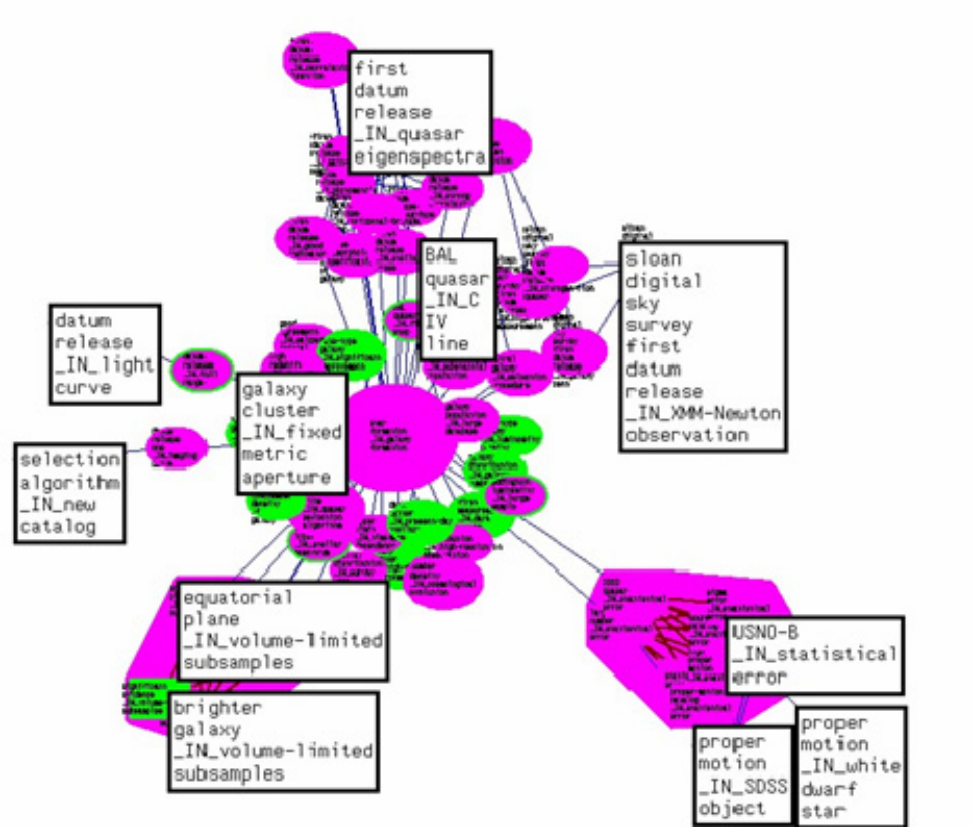

Figure 6. Atom graph of terms in abstracts referring to DR1. Wrapped atoms correspond to association rules of larger support.

Figure 6 shows the atom graph on DR1. The atom graph is centered on "*galaxy formation"* which means that the term *"galaxy formation"* is the cluster label and the term *"star formation"* is a member in the cluster that is the most highly connected to terms outside the cluster. However the atoms that explain the most frequent association rules are marginal as shown in the figure below. There are two atoms that are wrapped around "*volume limited subsamples*" and "*statistical error*" (lower left in the graph). As expected, atoms respect association rules and it is possible to map most of them to a formal concept.

4.3.2. Association rules and atom graphs based on Data Release 2

For the second data release, we obtained five item sets with high support.(18%)

{local universe, stellar mass} => {star-forming galaxy}
{local universe, star-forming galaxy} => {stellar mass}
{local universe} => {stellar mass}
{local universe} => {star-forming galaxy}

Figure 7. Top-5 association rules with high support for DR2

The atom graph is centered on *star-forming galaxy* which is embedded in a larger atom labeled *active nucleus.*

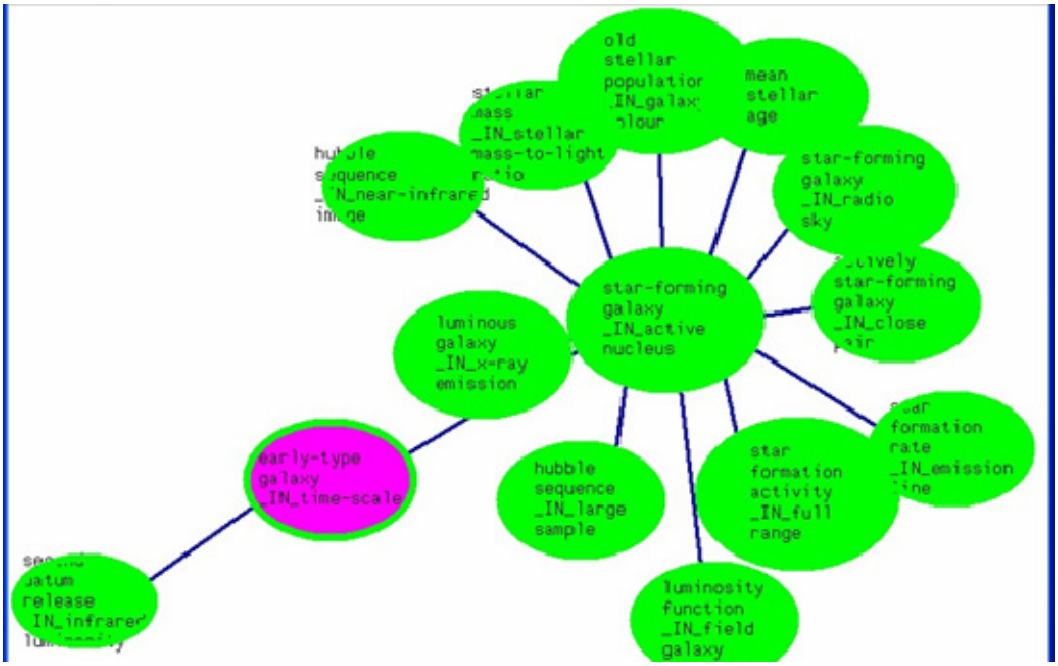

Figure 8. Atom graph obtained on DR2.

Note that the biggest largest atom *"full range"* is linked to the central atom *active nucleus* through the term "*star forming galaxy"*. "*Full range*" atom contains all terms involved in the six association rules as shown in figure 9 below where this atom is unfolded.

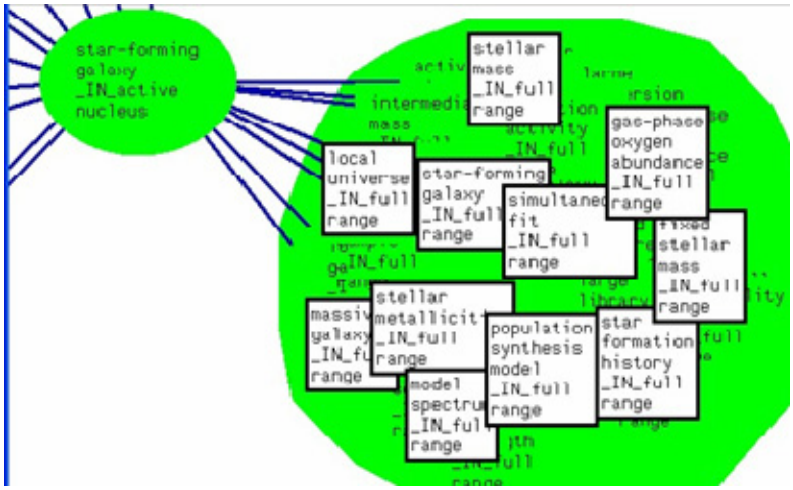

Figure 9. Unfolded version of the most central atom "full range" for DR2.

4.3.3. The general association graph revisited

Behind each association in graph of figure 3, it is possible to find an atom of terms that point out potential frequent item sets. By way of example, the atom graph in DR2 gives the context of association between "*star formation*" and "*intermediate mass galaxies*". The atom shows that all terms co-occur almost systematically together. As shown in figure 3, in terms of the color intensity of these vertices, this association appears to be significant in recent releases. However we can see that this relation was present since the DR2 and gave rise to the biggest atom extracted from this release, even though this atom is not central. In the following data release, this atom is not more frequent than the biggest one but it takes a more central position in the atom graph.

5. Discussion

A comprehensive analysis of the results shown here is beyond the scope of this conference and requires domain knowledge background. Our research is still ongoing on the impact of the different data releases on research on SDSS. This analysis could not be completed here owing to space limitations. However, we expect to have given some preliminary results which are promising in this direction. Also, we tried to clarify the relations between association graph analysis and frequent item sets mining on the base of atom graphs.

In order to proceed to a systematic evaluation of results from the combination of these methods, we are implementing a system of automatic summary of abstracts that supports an association or an atom. This will allow scientists to manage the two components of a formal concept easily, namely the intention that corresponds to the frequent item set and the extension made of documents that form the support of the frequent item set.

Acknowledgements

The work reported here is in part supported by the National Science Foundation under Grant No. 0612129.

References

Berry A., Kaba B., Nadif M., SanJuan E., Sigayret A. (2004). Classification et désarticulation de graphes de termes. In *JADT 2004.* Louvain-la-Neuve, Belgium, pp. 160-170.

Auber D., Chiricota Y., Jourdan F., Melancon G. (2003). Multiscale visualization of small world networks. In *IEEE Symposition on Information Visualisation, IEEE Computer Society*, pp. 75-81.

Berry M. W. (eds.). (2004). Survey of Text Mining. *Clustering, classification and retrieval.* Springer, NY, 244.

Chen C. (2008). An information-theoretic view of visual analytics. *IEEE Computer Graphics & Applications*. (Jan/Feb 2008).

Chen C. (2006). 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), pp. 359-377.

Chen C. (2004). Searching for intellectual turning points: Progressive Knowledge Domain Visualization. In *Proceedings of Natl. Acad. Sci. USA*, 101 (suppl.), pp. 5303-5310.

Ferrer-i-Cancho R., Sole R. V. (2001). The small world of human language. In *Proceedings of The Royal Society of London*. Series B, *Biological Sciences 268(1482)*, pp. 2261-2265.

Gamon M. (2006). Graph-Based text Representation for Novelty Detection. In *Proceedings of Workshop on TextGraphs*, at HLT-NAACL 2006, pp. 17-24.

Ganter B., Wille (1998). *Formal Concept Analysis: Mathematical Foundations*. Springer-Verlag, Berlin.

Ibekwe-SanJuan F., SanJuan E. (2004). Mining textual data through term variant clustering: the Termwatch system. In *Proceedings of the Conference "Recherche d'Information assistée par ordinateur",* (RIAO-04). Avignon, pp. 487-503.

SanJuan E. & Ibekwe-SanJuan F. (2006). Text Mining without document context. *Information Processing & Management*. Elsevier, 42(6), pp. 1532-1552.

Schiffrin R., Börner K. (2004). Mapping knowledge domains. *Publication of the National Academy of Science* (PNAS), 101(1), pp. 5183-5185.