Visualizing the LAK/EDM Literature Using Combined Concept and Rhetorical Sentence Extraction

Davide Taibi[1], Ágnes Sándor[2], Duygu Simsek[3],
Simon Buckingham Shum[3], Anna DeLiddo[3], Rebecca Ferguson[3]

[1]Institute for Educational Technologies
Italian National Research Council
Via Ugo La Malfa 153
90146 Palermo, Italy
davide.taibi@itd.cnr.it

[2] Parsing & Semantics Group
Xerox Research Centre Europe
6 Chemin de Maupertuis
F-38240 Meylan, France
agnes.sandor@xrce.xerox.com

[3] The Open University
Knowledge Media Institute &
Institute of Educational Technology
Milton Keynes, MK7 6AA, UK
firstname.lastname@open.ac.uk

ABSTRACT

Scientific communication demands more than the mere listing of empirical findings or assertion of beliefs. Arguments must be constructed to motivate problems, expose weaknesses, justify higher-order concepts, and support claims to be advancing the field. Researchers learn to signal clearly in their writing when they are making such moves, and the progress of natural language processing technology has made it possible to combine conventional concept extraction with rhetorical analysis that detects these moves. To demonstrate the potential of this technology, this short paper documents preliminary analyses of the dataset published by the Society for Learning Analytics, comprising the full texts from primary conferences and journals in Learning Analytics and Knowledge (LAK) and Educational Data Mining (EDM). We document the steps taken to analyse the papers thematically using Edge Betweenness Clustering, combined with sentence extraction using the Xerox Incremental Parser's rhetorical analysis, which detects the linguistic forms used by authors to signal argumentative discourse moves. Initial results indicate that the refined subset derived from more complex concept extraction and rhetorically significant sentences, yields additional relevant clusters. Finally, we illustrate how the results of this analysis can be rendered as a visual analytics dashboard.

Categories and Subject Descriptors

K.3.1 [**Computers and Education**]: Computer Uses in Education

General Terms

Design

Keywords

Learning Analytics, Corpus Analysis, Scientific Rhetoric, Visualization, Network Analysis, Natural Language Processing

1. INTRODUCTION AND MOTIVATION

Our overall aims are to provide users automatically with suggestions about similar papers, about connections between papers, and to present these similarities and connections in ways that are both meaningful and searchable.

In order to achieve this, we integrated three different approaches to linking and analysing a specific dataset of scientific papers (see section 2). These approaches were:

1. network analysis
2. rhetorical analysis
3. visualization of the results

Network analysis yields sets of related papers based on statistical corpus processing (Section 3). In order to improve the precision of information about the content of the connections among the papers, we carried out semantic and rhetorical analysis (Section 4). On the one hand, we extracted similar concepts in order to provide topical similarity indicators (Section 4.1) and, on the other hand, we extracted salient sentences that indicate the main research topics of these papers (Section 4.2). We repeated the statistical analysis of this reduced list of concepts, and of the reduced list of salient sentences. At the end of this paper, we present the design and implementation of the first prototype of an analytics dashboard (Section 5), which is designed to summarize results of the socio-semantic-rhetorical analysis in a way that users will find both meaningful and easy to explore.

2. THE LAK DATASET

We selected the LAK Dataset[1] published by the Society for Learning Analytics Research (SoLAR[2]), which provides machine-readable plain-text versions of the *Learning Analytics and Knowledge (LAK)* conference proceedings and a journal special issue related to learning analytics, and of the *Educational Data Mining (EDM)* conferences and journal.

The corpus was extracted using the SPARQL endpoint of the LAK dataset. The corpus comprised the following:

- 24 papers presented at the LAK2011 conference
- 42 papers presented at the LAK2012 conference
- 10 papers from the journal of Educational Technology and Society special issue on learning analytics
- 31 papers presented at the EDM2008 conference
- 32 papers presented at the EDM2009 conference
- 64 papers presented at the EDM2010 conference
- 61 papers presented at the EDM2011 conference
- 52 papers presented at the EDM2012 conference

For each resource, the title, description and keywords properties were used to feed the data mining processes employed in our analysis. At the end of this initial process, a relational database was used to store 305 papers, 599 authors, 448 distinct keywords. After this preliminary phase the entire LAK Dataset

[1] LAK Dataset: http://www.solaresearch.org/resources/lak-dataset

Published by SoLAR and made available to the LAK Data challenge of the 3rd International Conference on Learning Analytics and Knowledge (http://lakconference.org)

[2] http://www.solaresearch.org

was analyzed by using the Xerox Incremental Parser (XIP) [1] for concept extraction and rhetorical analysis, a total of 305 papers, from which XIP extracted 7,847 sentences and 40,163 concepts.

3. STATISTICAL ANALYSIS

A preliminary analysis reported the most-used keywords, the most frequently occurring authors and the most-referenced papers. A second phase of analysis was then carried out using the data-mining tool, *RapidMiner* [2].

3.1 Statistical Data from *RapidMiner*

A three-step process was developed in order to analyze the corpus using the data-mining tool, *RapidMiner*:

- **Process documents from file**: this module generates word vectors from the text files.
- **Select attributes**: This allows users to select the attributes to be considered by the analysis. In our case, a threshold was set in order to eliminate less important elements in the word vectors.
- **Data to similarity**: This module was used to calculate a similarity index for the conference papers based on Cosine similarity.

The first block '*Process Documents from file*' is made up of the following steps:

- **Tokenize**: This operator splits the text of a document into a sequence of tokens.
- **Replace token**: This operator is used to replace tokens, for instance in cases where words are misspelled.
- **Filter tokens (by length)**: This operator filters tokens based on their length. In our case, all the words with fewer than three characters were removed.
- **Filter stopwords (English)**: This operator filters English stopwords from a document by removing every token that is the same as a stopword from the built-in stopword list.
- **Stem (Snowball)**: This operator stems words by applying stemming using the Snowball tool.[3]

At the end of the main process, the '*Data to Similarity*' step returns two results:

a) The list of the most relevant words (stemmed version) used in the entire corpus

b) The measured similarity index between the papers that make up the corpus.

We employed the similarity relationships between papers to build a network of papers. In this network each node represents a paper, and an edge between two paper is created if the similarity value of a pair of papers overcome a threshold of 0.3.

3.2 Analysing the Network of Papers

The network of papers was then analysed with the *yEd* tool[4] in order to extract clusters of documents using the algorithm for natural clusters "based on Edge Betweenness Clustering proposed by Girvan and Newman" [3]. This algorithm has been successfully used in Network Analysis to study communities and their aggregations [4]. The yEd tool allows users to balance quality and speed of the cluster algorithm by the use of a slider. When the quality is set at the highest value, the Girvan and Newman algorithm is used in its normal form. At the opposite end, the lowest quality value produces the fastest running time. In this case it executes a local betweenness calculation following Gregory's algorithm [5]. When a mid value is chosen for quality and speed, the fast betweenness approximation of Brandes and Pich [6] is applied. In this case, less accurate clustering is balanced by a lower execution time.

The clusters created with *yEd* have the following properties:

- each node (paper) is a member of exactly one cluster
- each node shares many edges with other members of its cluster, where edges represent the connection between a pair of papers if their similarity values is more than a threshold value (0.3 in our experiment).
- each node shares few or no edges with nodes of other clusters

Figure 1 shows a visualization of the primary clusters. Some of the clusters did seem to have thematic coherence, while others were harder to label:

- Cluster 1: collaborative, learning, social
- Cluster 2: skills, model, slip, guess, parameters
- Cluster 3: causality, variables, model, construct
- Cluster 4: question, fit, grain, school, skill
- Cluster 5: translating, sentences, grinder, corpus

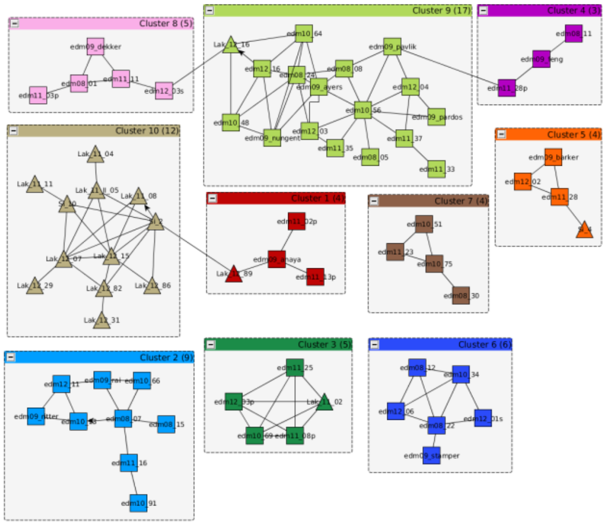


Figure 1: Results of initial LAK paper clustering analysis

The complete list of the papers belonging to the clusters has been reported in the web page[5] associated to this work.

This analysis was word-driven and not concept-driven. The next step was to try and refine this by distilling (1) a richer set of concepts, and (2) a more salient subset of sentences.

4. SEMANTIC ANALYSIS

In order to go beyond full-text statistical analysis and find connections between papers at the level of the claims they make, we processed the corpus using the Xerox Incremental Parser

[3] http://snowball.tartarus.org

[4] http://www.yworks.com

[5] http://www.pa.itd.cnr.it/lak-data-challenge.html

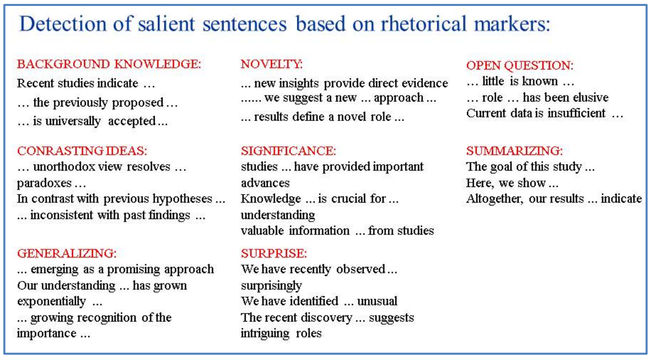
(XIP) [1] for extracting concepts and rhetorically salient sentences [7].

4.1 Concept Extraction

The basic module of XIP performs morphosyntactic analysis, part-of-speech tagging, constituent analysis and dependency extraction on free text. Since we define concepts as simple or compound noun phrases, they can be identified using general morphosyntactic analysis. Examples of extracted concepts are *analytics, learning analytics, social learning analytics* and *social network analytics.*

4.2 Rhetorical Analysis

Scientific research does not consist in providing a list of facts, but in the construction of narrative and argumentation around facts. In articles, researchers make hypotheses, support, refute, reconsider, confirm, and build on previous ideas in order to support their ideas and findings. The aim of rhetorical analysis is to detect where authors signal that they are making such moves. This analysis builds on the widely studied feature of research articles that, besides their well-defined standard structure (title, abstract, keywords, often IMRAD body structure) rhetorical moves emphasize articles' contribution to the state of the art, and the research problems they address. In previous work [7] we described a list of rhetorical moves that characterize such salient messages, together with the extraction methodology. Figure 2 lists the detected rhetorical moves (in caps) together with examples of expressions that mark them.



Detection of salient sentences based on rhetorical markers:

BACKGROUND KNOWLEDGE:
Recent studies indicate …
… the previously proposed …
… is universally accepted …

NOVELTY:
... new insights provide direct evidence
...... we suggest a new ... approach ...
... results define a novel role ...

OPEN QUESTION:
… little is known …
… role … has been elusive
Current data is insufficient …

CONRASTING IDEAS:
… unorthodox view resolves …
paradoxes …
In contrast with previous hypotheses ...
... inconsistent with past findings ...

SIGNIFICANCE:
studies ... have provided important advances
Knowledge ... is crucial for ... understanding
valuable information ... from studies

SUMMARIZING:
The goal of this study ...
Here, we show ...
Altogether, our results ... indicate

GENERALIZING:
... emerging as a promising approach
Our understanding ... has grown exponentially ...
... growing recognition of the importance ...

SURPRISE:
We have recently observed ...
surprisingly
We have identified ... unusual
The recent discovery ... suggests intriguing roles

Figure 2: Rhetorical moves (in capital red letters) followed by some examples of expressions used to signify them in papers

Once the XIP concept extraction and rhetorical analysis were concluded we repeated the cluster analysis on the XIP-filtered lists of concepts and salient sentences. Thus our statistical analysis (described in Section 3.) of the LAK dataset has been conducted in three different ways:

- considering the full text of the articles
- considering only the salient sentences extracted by XIP
- considering only the concepts extracted by XIP

The comparison of the sets of papers yielded by the three approaches is still ongoing. At this stage we can only present some preliminary observations concerning pairs of similar papers yielded by the three kinds of input. The data obtained through this preliminary evaluation is reported in the web page[6] related to this work.

A basic observation concerns the distribution of the pairs of similar papers yielded by the three methods. According to the expectations, the most similarity pairs have been yielded by taking into account the full text only in both the LAK and the EDM collection. There are considerable overlaps among the three methods, and there are cases when just one method yields similarity pairs. In subsequent evaluations we aim at evaluating these various cases. As a first step towards a more complete evaluation, we have selected some pairs of papers and checked their similarity according to some independent similarity indicators. We have found that our statistical method is coherent with independent similarity indicators in case of high similarity scores and that in these cases, similarity is found with and without XIP-extracted text. This indicates the validity of our statistical method in these cases for finding related papers. In the case where no independent similarity indicator could be found, but we do have XIP-based similarity pairs, we looked for related key claims or findings in the pairs of papers[7]. In the cases where the similarity score between the two papers was high we did find such interesting related claims in the two papers. However, in cases where the similarity measure is low, we did not find any related claims. This indicates that we might want to define a threshold score. The details of the preliminary tests are reported in the web page.

5. XIP DASHBOARD

The XIP Dashboard was designed to provide visual analytics from XIP output in order to help readers assess the current state of the art in terms of trends, patterns, gaps and connections in the LAK and EDM literature. The dashboard also draws attention to candidate patterns of potential significance within the dataset:

- the occurrence of domain concepts in different metadiscourse contexts (e.g. *effective tutoring dialogue* in sentences classified as *contrast*).
- trends over time (e.g. the development of an idea)
- trends within and differences between research communities, as reflected in their publications.

5.1 Implementation

All the papers in the LAK dataset were analyzed using XIP. The output files of the XIP analysis, one per paper, were then imported into a MySQL database, and the user interface was implemented using PHP and JavaScript, making use of Google Chart Tools for the interactive visualizations.[8]

5.2 User Interface

The dashboard consists of three sections, each showing different analytical results in different types of chart.

Section one of the dashboard shows two line charts, representing the LAK and the EDM conferences respectively. Each line chart shows the distribution of the number of salient sentences over time and by rhetorical marker type (see Figure 2 for a list of the types of rhetorical markers). Each coloured line in these line charts indicates how many sentences of a specific rhetorical type were extracted, and how this number changed by year (Figure 3 shows the line chart for the EDM conference).

[6] http://www.pa.itd.cnr.it/lak-data-challenge.html

[7] The related claims have been searched by reading the pairs of sentences. Our long-term goal is to provide the related claims automatically.

[8] https://developers.google.com/chart

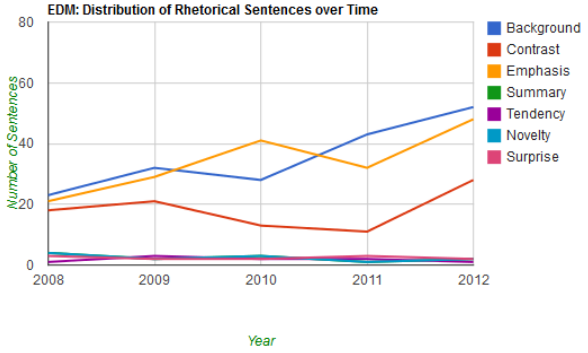


Figure 3: Rhetorical sentences graphed by year, for EDM

The second section of the dashboard (Figure 4) allows users to select a combination of the extracted concepts, in order to visualize the occurrence of these concepts in papers within any or all research communities represented in the corpus– that is to say across the whole LAK dataset (EDM plus LAK conference).

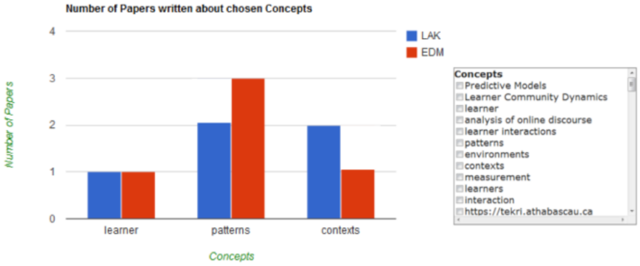


Figure 4: Number of papers with rhetorically extracted sentences containing user-selected concepts

The third dashboard section consists of a bubble chart that displays the occurrence of papers within the entire dataset, filtered by user-selected concepts (Figure 5). This visualization can be restricted to display just the LAK or the EDM conference. In Figure 5, each bubble represents a concept that has been selected by the user. This is associated with a specific number of papers and sentences in which that concept has been detected. The colour saturation of each bubble (expressed by the color spectrum shown at the top) represents the 'density' of the chosen concept as defined by the number of XIP-extracted sentences in which the concept occurs. The darker the colour, the greater the density.

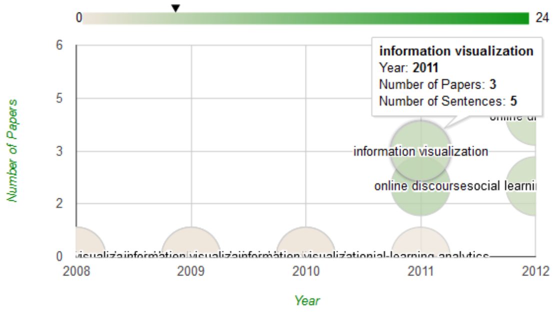


Figure 5: Concept 'density' within XIP sentences, by year and number of papers

When a concept bubble is selected (Figure 6), a pie chart pops up representing the relative distribution of the rhetorical types for that bubble (that is to say for that concept, and across the papers and sentences in which the concept has been detected).

6. SUMMARY

This short paper has summarised an approach to conducting 'analytics on Learning Analytics'. The LAK Dataset comprising LAK and EDM literature has been analyzed in order to identify clusters of papers dealing with similar topics (conceptual clustering), and in order to identify key contributions of papers in terms of the claims authors make, as signalled by rhetorical patterns. Our preliminary tests are promising, but more thorough testing is needed to validate the method. Finally, we showed how the results of this analysis are beginning to be visualized using an analytics dashboard. All the secondary datasets produced have been published as open data, for further research.

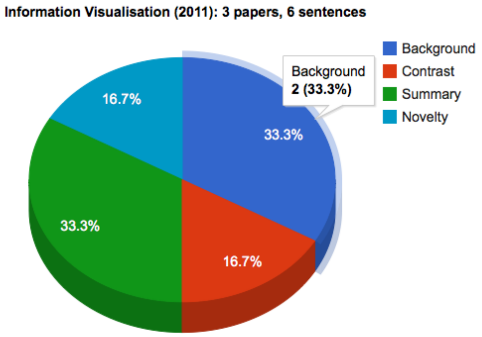


Figure 6: Distribution of rhetorical types in XIP-classified sentences within a selected concept bubble

In the longer term, the aim of this research is to provide users with automatic suggestions about similar papers and about connections between papers, and to present these similarities and connections in ways that are both meaningful and searchable for the users. Future steps will validate the outputs from these analyses with researchers, and test the usability of the dashboard with different end-users (e.g. researchers, educators, students).

7. REFERENCES

[1] Salah Aït-Mokhtar, Jean-Pierre Chanod, and Claude Roux. (2002). Robustness beyond shallowness: incremental dependency parsing. *Natural Language Engineering*, 8(2/3):121-144.

[2] Jungermann, F. (2009). Information extraction with RapidMiner. In *Proceedings of the GSCL Symposium' Sprachtechnologie und eHumanities'*. W. Hoeppner, ed.

[3] Girvan M. and Newman. M. E. J. 2002. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*. 99, 12, 7821-7826.

[4] Newman MEJ: Detecting Community Structure in Networks. Eur Phys J B 2004, 38:321-330.

[5] Gregory, S.: Local Betweenness for Finding Communities in Networks. *Technical Report*, University of Bristol (2008).

[6] Brandes, U., Pich, C., Centrality Estimation in Large Networks. *Intl. Journal of Bifurcation and Chaos in Applied Sciences and Engineering* 17(7) 2303–2318

[7] Ágnes Sándor. (2007). Modeling metadiscourse conveying the author's rhetorical strategy in biomedical research abstracts. *Revue Française de Linguistique Appliquée* 200(2): 97-109