Ontology Learning to Analyze Research Trends in Learning Analytics Publications

Amal Zouaq
Department of Mathematics and Computer Science
Royal Military College of Canada
Kingston, ON, Canada
+1 613 541 6000, Ext. 6478
amal.zouaq@rmc.ca

Srećko Joksimović
School of Interactive Arts and Technologies
Simon Fraser University
Surrey, BC, Canada
+1 778 782 7474
sjoksimo@sfu.ca

Dragan Gašević
School of Computing and Information Systems
Athabasca University
Athabasca, AB, Canada
+1 604 569 8515
dgasevic@acm.org

ABSTRACT
In this paper, we show how ontology learning tools can be used to reveal (i) the central research topics that are tackled in the published literature on learning analytics and educational data mining; and (ii)relationships between these research topics and iii) (dis)similarities between learning analytics and educational data mining.

Categories and Subject Descriptors
I.2.7 [**Artificial Intelligence**]: Natural Language Processing; G.2.2 [**Discrete Mathematics**]: Graph Theory

General Terms
Algorithms, Measurement, Experimentation

Keywords
Ontology learning, deep parsing, filtering, information retrieval, ranking algorithms, graph theoretic statistics

1. INTRODUCTION
Learning analytics is a new research discipline. Although it attracted a considerable amount of attention in educational research and practice, debate is still very active about the scope of the discipline. The definition of learning analytics offered by the Society for Learning Analytics Research [7], which is commonly used in the literature to date, gives a general framework for the main tasks learning analytics are about. However, given the youth of the discipline, there are generally two open questions:

- What are the central research topics that are tackled in the published literature?
- What are the relationships between the central research topics?
- What are similarities and differences between learning analytics and educational data mining?

To address the above questions, we aimed to analyze systematically textual content available in the LAK Challenge data set. In particular, we used a state-of-the-art ontology learning tool, OntoCmaps, that enabled the automatic (i) parsing of textual content, (ii) creation of conceptual maps based on the extracted concepts and relationships, and (iii) filtering/ranking of the most important concepts and relationships based on measures of information retrieval, graph theory, and voting theory. The concept extraction and their filtering/ranking was done (i) for each edition of the two conferences and the journal special issue (from the LAK 2013 Challenge dataset)individually (i.e., LAK 2011-2012, EDM 2008-2013, and LAK ET&S special issue) to see the emerging trends through the years; and (ii) by creating two subsets – one for the papers presented at the LAK conference editions and another one for the papers presented at the EDM conference editions – in order to compare the two conferences based on concepts and relationships gauged as most important. We also performed analysis based on (a) paper abstracts only and (b) main body of text of the papers.

In this short report, we first describe the data analysis pipeline. This is followed by a very brief discussion of a small fragment of the results we obtained in our analysis. The complete results in the CSV format are available at [8].

2. DATA ANALYSIS PIPELINE
The data analysis relies on our ontology learning tool, OntoCmaps[10]. Ontology learning from text is a multi-layer knowledge extraction task that targets the following components:

Terms and concepts: The first step consists in identifying candidate expressions in texts. These expressions are then ranked using some kind of measure (statistical metrics, graph-based metrics, etc.) to extract those that are relevant for the domain. These filtered relevant expressions are then considered "concepts" in the ontology learning community.

Taxonomy: This step identifies "is-a" links in texts, generally using patterns indicating a taxonomical link in text such as Hearst's patterns[11], or using the inner structure of multiword expressions. For example, a "carnivorous plant" can be considered a "plant" just by looking at the syntactic structure "Adjective noun" of the expression.

Conceptual relationships: This step uses various techniques (patterns, machine learning, etc.) to identify any kind of transversal relations, with a domain and range.

Axioms: Finally, axioms here mean defined classes, or rules from texts.

OntoCmaps requires a domain corpus as input. As such, LAK and EDM proceedings (the LAK dataset [13]) were an appropriate set of texts to test the ontology learning process. OntoCmaps relies on three main phases to learn a domain ontology: 1) the extraction phase that performs a deep semantic analysis based on dependency patterns; 2) the integration phase that builds concept maps, which are composed of terms and labeled relationships, and uses basic disambiguation techniques. These concept maps form a graph; and finally 3) the filtering phase where various metrics rank the items (terms and relationships) in concept maps.

2.1 The Extraction Phase
In the **extraction phase**, OntoCmapsis based on a hierarchy of syntactic patterns. Each pattern describes a set of syntactic rela-

tionships that permit the extraction of a "semantic representation". OntoCmaps does not rely on any predefined domain knowledge. It uses two NLP tools to obtain the syntactic representations: the Stanford Parser along with its dependency module [2] and the Stanford parts-of-speech (POS) Tagger [6]. Given a sentence, the Stanford parser generates syntactic dependency relations between each pair of related words of a sentence. The POS Tagger identifies words' parts-of-speech. Based on these two inputs, OntoCmaps creates a pattern syntactic format that enriches words in each dependency relation with their parts-of-speech. This enriched representation is then used as input to a pattern recognition task. A recognized pattern fires a rule that applies various transformations on the syntactic representation to obtain a "semantic representation", in the form of expressions, triples or sets of triples. The patterns are divided into conceptual patterns and hierarchical patterns. Hierarchical patterns concentrate on the extraction of taxonomical links, following the work of [11], but based on the dependency formalism. Conceptual patterns identify the main structures of the language that can be transformed into triples useful for the extraction of conceptual relations. They are organized into a hierarchy from most-detailed patterns (containing the biggest number of dependency relationships) to least detailed. The extraction phase targets deeper levels of the hierarchy first to avoid extracting too abstract or incomplete representations. For instance, if the pattern "nsubj-dobj-xcomp" exists in text, the extractor should fire it instead of firing one of its higher-level counterparts "nsubj-dobj" and "nsubj-xcomp"which contain only a subset of the syntactic relationships of interest. If a pattern is instantiated, then all its parents in the hierarchy are disregarded.

2.2 The Integration Phase

In this **integration phase**, all the extracted relationships are gathered into concept maps. Some basic term disambiguation tasks are performed at this level mainly: i) lemmatization which considers singular, plural and other forms of the same terms or relationships as referring to a single concept or relationship; ii) basic synonym detection based on abbreviation relations that are generated by the Stanford parser and iii) a kind of co-reference resolution phase that is built in some of the patterns, and that allows for the creation of semantic links between terms in a sentence, even if not direct dependency links existed in the original dependency representation. For example, in the sentence: *carnivorous plants are organisms which eat insects*, the co-reference resolution creates a relation "*eat*" between the term "*carnivorous plants*" and the term "*insects*" while the grammatical representation links the term "*plants*" to the term "*insects*".

All these operations result in concept maps around various terms. For example, if there were a number of statements around the term "*carnivorous plants*" in texts, it is likely that a concept map around "*carnivorous plants*" will be created. This process is repeated for all identified terms and relationships and results in an aggregation of concept maps through links between various concept maps, thus constituting a graph, with terms representing nodes, and relationships representing edges.

2.3 The Filtering Phase

The third and last phase for learning the domain ontology is the **filtering phase**, which aims at ranking the items in concept maps (domain terms, taxonomical links, and conceptual links).

2.3.1 Concept Filtering

A number of metrics from graph theory and from information retrieval are used to identify relevant terms. Graph-based metrics were computed using the JUNG framework [3]. These metrics include:

- The Degree centrality of a node which identifies the number of edges from and to a given node.
- The Betweenness centrality, which assigns each node a value that is derived from the number of shortest paths that pass through it;
- The HITS algorithm which ranks nodes according to the importance of hubs and authorities [5]. This resulted in two measures Hits-Hubs and Hits-Authority;
- The PageRank of a node [1];
- We also computed standard information retrieval metrics, mainly term frequency (TF) and TF-IDF.

Finally, using the graph-based metrics, we defined a number of voting schemes with the aim of improving the precision of filtering. All the VS relied on three metrics that were identified as being among the best metrics in previous experiments [10][11]: Degree, Betweenness and HITS-Hubs. The VS include:

- The majority voting scheme, which recognizes a term as an important one if it is chosen by at least k metrics out of n with k>n/2.
- Borda Count Voting Scheme: This method assigns a "rank" to each candidate. A candidate who is ranked first receive n points (n=size of the domain terms to be ranked), second n-1, third n-2 and so on. The "score" of a term for all metrics is equal to the sum of the points obtained by the term in each metric.
- Nauru Voting Scheme: The Nauru voting scheme is based on the sum of the inverted rank of each term in each metric. It is used to put more emphasis on higher ranks.

Table 1 shows the top ranked concepts based on the majority voting scheme. All the base metrics (Betweenness, PageRank, Degree, etc.) and voting schemes have been computed and can be found at [8]. The Web site [8] also features a visualization of the extracted data based on the obtained concept maps. The visualization is performed per venue (EDM/LAK/ETS-SI), per corpus (only abstracts or main texts) and per year (2008-2012).

2.3.2 Relationship Filtering

Similarly, a number of metrics were used to identify important relationships.

The first measure consists of all the relationships that occur between important terms (determined through the voting schemes) as important relationships. This constitutes our voting schemes for relationships, which were based on the results of the majority voting scheme for concepts.

The second measure ranks relationships based on Edge Betweenness centrality, which is a measure of the importance of edges based on the number of shortest paths which contain them.

The third measure is based on assigning frequencies of co-occurrence weights based on the Dice coefficient [9], a standard measure for semantic relatedness.

Table 2 shows an excerpt of the top ranked relationships based on the majority voting scheme. Contrary to standard named entity extractors, an important aspect of using ontology learning is the ability to extract relationships as well, thus, obtaining not only topics but also relationships (taxonomical and conceptual) between these topics. A better approach would mix the two approaches and combine topic extraction using named entity extractors, linked data semantic annotators and ontology learning.

Table 1.Top ranked concepts based on the majority voting scheme extracted the subsets of the LAK 2013 Challenge dataset

LAK (abstracts)	LAK (paper body)	EDM (abstracts)	EDM (paper body)
student (0.50)	student (0.75)	student (0.75)	student (0.75)
datum (0.45)	datum (0.20)	model (0.38)	model (0.23)
informal_learn (0.31)	learner (0.15)	datum (0.37)	datum (0.19)
learn (0.31)	course (0.15)	method (0.19)	skill (0.09)
teacher (0.29)	analysis (0.12)	paper (0.16)	problem (0.08)
model (0.27)	activity (0.11)	system (0.13)	result (0.06)
learning_analytics (0.26)	user (0.10)	result (0.12)	method (0.06)
learner (0.25)	tool (0.10)	approach (0.11)	parameter (0.05)
social_factor (0.21)	learn (0.09)	skill (0.08)	question (0.05)
social_learn (0.19)	analytics (0.07)	analysis (0.07)	performance (0.05)
effective_learn (0.19)	group (0.07)	intelligent_ tutoring_system(0.07)	system (0.05)
group_learn (0.17)	system (0.07)	behavior (0.07)	approach (0.04)
knowledge_ professional (0.17)	teacher (0.06)	tool (0.07)	example (0.04)
Lak (0.17)	instructor (0.06)	work (0.06)	feature (0.04)
knowledge (0.17)	network (0.06)	Researcher (0.06)	item (0.04)

Table 2.Top ranked relationships based on the majority voting scheme extracted the subsets of the LAK 2013 Challenge dataset. Each cell in the table contains a concept-relationship-concept triplet

LAK (abstracts)	LAK (paper body)	EDM (abstracts)	EDM (paper body)
learner–build–knowledge (1)	course–being recorded as well as to–student (1)	datum–mining–method (1)	model–fit–student (1)
datum–obtained from–learner (0.81)	datum–break ability to educate effectively–student (0.60)	method–linguistics in–paper (0.95)	datum–are collected far from–student (0.96)
learning_analytics–important step for–teachers_of_tomorrow (0.78)	system–addresses indivi-dually–student (0.45)	model–are trained over–datum (0.70)	skill–will have been covered by–student (0.67)
teachers_of_tomorrow–is a–teacher (0.77)	analysis–have since been moved as–student (0.37)	system–provides–student (0.61)	problem–assign for–student (0.67)
tool–incorporate functionality to access–datum (0.65)	network–impacting–student (0.31)	student–are represented by–model (0.56)	example–parameterization by–student (0.63)
model–can be used to inform–student (0.64)	process–finally should promote reflection on–instructor (0.29)	model–can detect–student (0.50)	question–were based–student (0.62)
datum–obtained from–instructor (0.62)	tool–identify–student (0.27)	datum–derived from–student (0.43)	student–provides useful evidence to–model (0.60)
learner–generating–datum (0.58)	datum–may be presented to–learner (0.25)	goal–has been investigated by–researcher (0.42)	step–requires–student (0.57)
student–accessing–online_discussion_forum (0.56)	activity–conducted by–user (0.25)	tutoring_system–is a–system (0.40)	performance–dependent upon–student (0.56)
model–can be used to inform–teacher (0.51)	group–will contain–student (0.25)	student–study with–intelligent_tutoring_system(0.39)	accuracy–varies across–student (0.48)
student–flock to–online_service (0.48)	environment–capture–datum (0.24)	skill–studied in–tutoring_system (0.38)	student–is guessing–result (0.48)
datum–are combined to calculate–likelihood_of_student (0.45)	model–highly accurate on–student (0.22)	intelligent_tutoring_system–are informed by–datum (0.32)	student–collect–datum (0.45)
instructor–guide–student (0.39)	average–miss–student (0.21)	analysis–reveals–unexpected_result (0.30)	word–uttered by–student (0.44)
learn–integral to–success_of_community (0.37)	role–are imposed on–student(0.21)	unexpected_result–is a–result (0.30)	datum–were used to build–model (0.44)
likelihood_of_student–is related to–student (0.36)	information–useful for–student (0.20)	collaborative–learning–interactions_of_student (0.29)	skill–are included in–model (0.41)

We can also notice that we were not always successful in extracting meaningful relationships labels from this corpus. One possible explanation is the type of texts (publications) and the amount of noise in these texts. In fact, OntoCmaps is made to run on clean plain sentences that describe a domain of interest and define it. Parts of research papers such as figure captions, formulas, and references represent noise for OntoCmaps. Additional cleaning of the input texts would be necessary. However, even when the labels were not meaningful, the existence of a link between two concepts (unlabeled relationship) was shedding some light on the domain (see Section 3).

3. FINDINGS

In this section, we present only results of the 15-top ranked concepts and relationships according to the Majority Voting Scheme (Betweenness, Degree, and Hits-Hub) as shown in Tables 1-2 (N.B. As can be noticed in the tables, the majority of the terms are lemmatized, that is, we show only their lemma or root. For example,*informal_learn* for *informal learning* or *datum* for *data*. In few cases, such as *learning_analytics*, the lemmatizer returned the expression itself). First, we could not possible include all the results of all the metrics we calculated in our experiment (those results are available at [8]). Second, we selected the metrics which were proven to be most accurate in our previous research [10], [11]. Finally, it should be noted that the purpose of our experiment here was not to evaluate the effectiveness of individual metrics, but rather to experiment if ontology learning technology can shed some light on the questions posed in the introduction of relevance to the LAK 2013 Data Challenge.

Concepts reported in Table 1 reveal that papers of both the LAK and EDM conferences have students, data and models as shared concepts. However, it is clear that LAK papers also focus on teachers/instructors, informal learning, and social, networked, and group learning. On the other hand, EDM papers focus on (data mining) methods and approaches, intelligent tutoring systems, features (extraction), and various types of parameters.

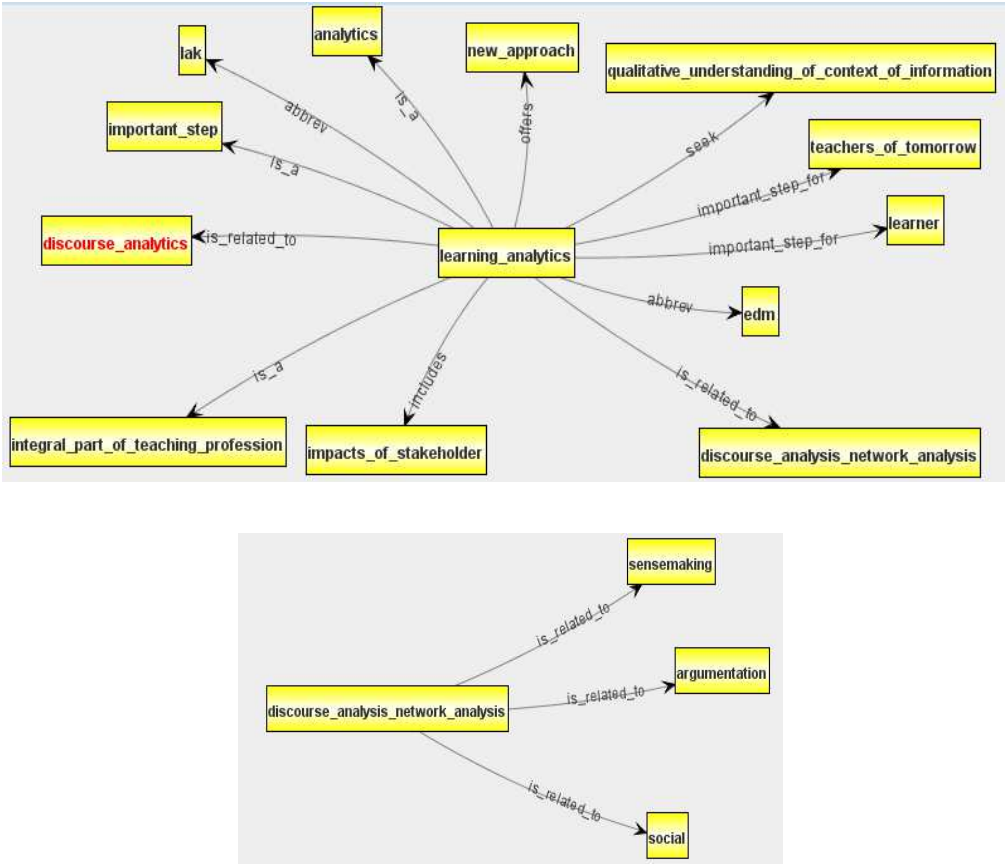

Figure 1.Two conceptual maps extracted from the abstracts of the papers presented at the LAK conference

Relationships reported in Table 2 further corroborate the observation that the LAK papers are more focused on teachers in order to empower them with learning analytics and to help them guide students. Moreover, there is an emphasis on (promoting) reflection of both students and instructors. Various aspects of social learning such as role playing and impact of communities appear to be highly popular topics in the LAK papers. On the other hand, EDM papers are much more focused on intelligent tutoring systems, accuracy of different types of (predictive) models, and revealing unexpected patterns. Certainly, focus on data is shared by both the LAK and EDM communities, but LAK also seems to be focused on data collected by and for instructors, not only for students. This probably indicates a trend that the LAK community has so far acknowledged the role of instructors in the learning process and aimed at supporting them as much as learners. The EDM community has however focused more on measuring and predicting specific types of skills. This is consistent with their focus on intelligent tutoring systems in which automated assessment of learners' skills is of paramount importance.

Finally, we were also able to visualize the extracted conceptual graphs. In Figure 1, we show the relationships of concept *learning analytics* as extracted from the abstracts of the papers presented at

the LAK conference. This figure further corroborates earlier observations by indicating that learning analytics is an integral part of teaching profession, is an important step for teachers of tomorrow and learners, and offers a new approach. This figure reveals also the nature of learning analytics to promote qualitative understanding of context of information. Learning analytics is also (strongly) related to discourse analytics, which seems to be consistent with the strong emphasis of learning analytics on social learning and which is further confirmed by extracted relationships of discourse learning analytics with sense-making, argumentation and social, all of which are types of skills recognized as important for the modern society.

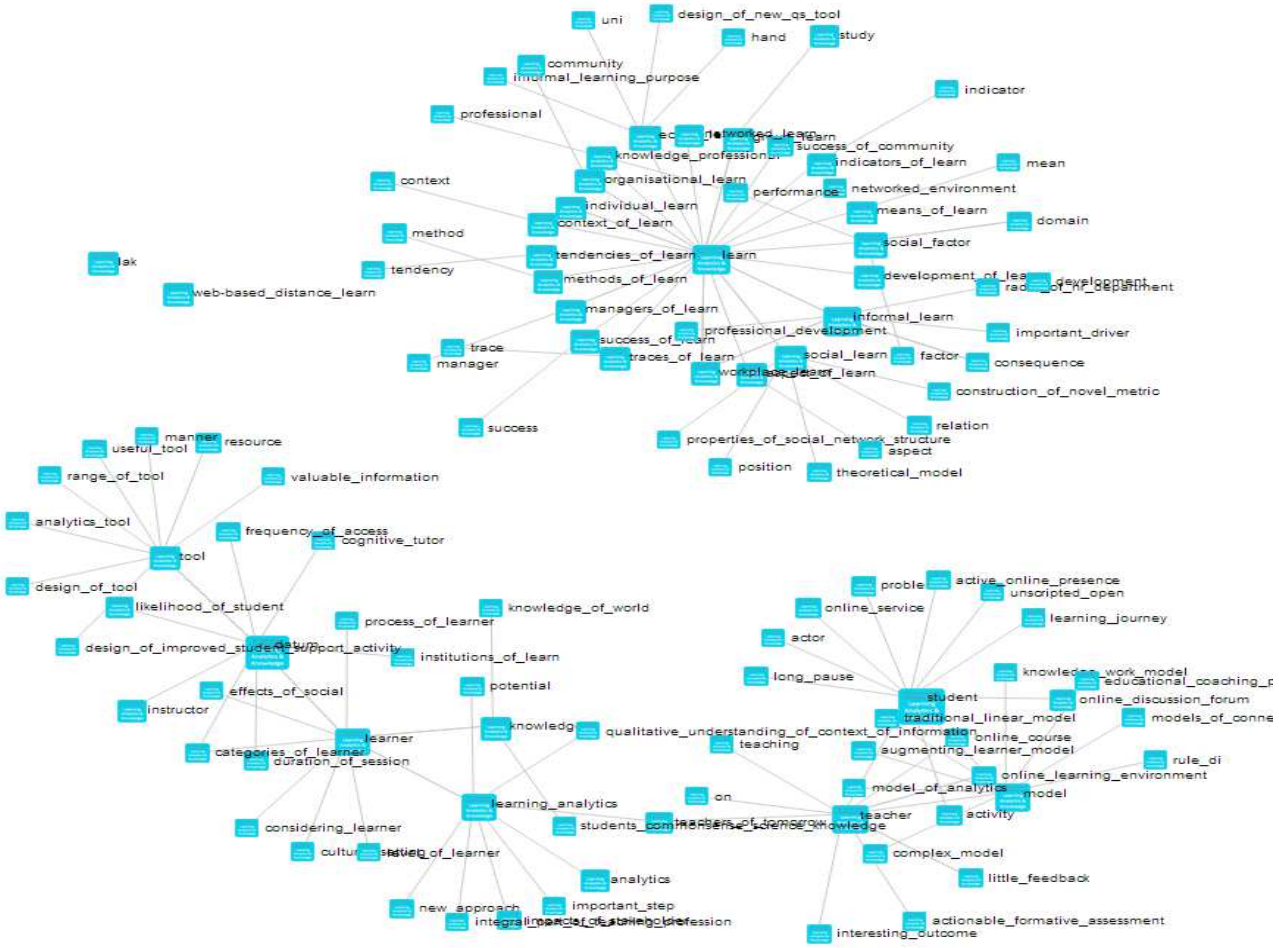

Figure 2. Visualization of top 30 ranked concepts based on the majority voting scheme extracted from the abstracts of the LAK 2013 Challenge dataset.

In future work, we plan to analyze further the research trends over the years for the LAK and EDM communities. Another of our goals is to compare the extractions of an ontology learning system such as OntoCmaps with Linked data Semantic Annotators such as DBPedia Spotlight[1] or Alchemy[2].

4. CONCLUSION

Funnily, our text analysis tool inferred that *EDM is an abbreviation of learning analytics*. This probably comes from the open debate reflected in the analyzed papers about the relationships between learning analytics and educational data mining. We hope that this paper sheds some light on the (dis)similarities of the two areas. We also hope that our analysis of the LAK 2013 Data Challenge dataset with the ontology learning tools indicated a high potential of this type of analytics to help the research community of new research discipline define itself and relationships with closest communities. More interesting results are available on our website [8]. For example, those results allow for (i) comparing results of different concept/relationship measures and (ii) chronological trends emerging throughout the years of individual editions of both the conferences. An example of one of the visualizations available at [8] is presented in Figure 2.

Of course, ontology learning tools are not perfectly accurate, and thus, few "strange" concepts and relationships are shown in our tables. An opportunity is however in combining such ontology learning tools as starting points of the concept map development of the learning analytics domain, which can then be refined through crowd sourcing (e.g., in a Wiki-like manner).

5. REFERENCES

[1] Brin, S. & Page, L. (1998). The anatomy of a large-scale hyper-textual web search engine, Stanford University.

[2] De Marneffe, M-C, MacCartney, B. and Manning. C.D. (2006). Generating Typed Dependency Parses from Phrase Structure Parses. In Proc. of LREC, pp. 449-454, ELRA.

[1]https://github.com/dbpedia-spotlight/dbpedia-spotlight/

[2]http://www.alchemyapi.com/

[3] JUNG (2013). Last retrieved from http://jung.sourceforge.net/

[4] Klein, D. and Manning, C.D. (2003). Accurate Unlexicalized Parsing. Proc. of the 41st Meeting of the Association for Computational Linguistics, pp. 423-430.

[5] Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment, Journal of the ACM 46(5): 604-632, ACM.

[6] Toutanova, K., Klein, D., Manning, C.D. & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network, In Proc. of HLT-NAACL, pp. 252-259.

[7] http://www.solaresearch.org/mission/about/

[8] http://lakchallenge.co.nf

[9] Van Rijsbergen, CornelisJoost (1979). *Information Retrieval*. London: Butterworths. ISBN 3-642-12274-4.

[10] Zouaq, A., Gasevic, D. and Hatala, M. (2011). Towards Open Ontology Learning and Filtering, Information Systems, 36(7): 1064–1081.

[11] Zouaq, A., Gasevic, D. and Hatala, M. (2012a). Voting Theory for Concept Detection. The 9th Extended Semantic Web Conference 2012 (ESWC 2012), pp. 315-329.

[12] Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In Proc.14th Conference on Computational Linguistics – Vol. 2 (COLING '92), 539-545.

[13] Taibi, D., Dietze, S., Fostering analytics on learning analytics research: the LAK dataset, Technical Report, 03/2013, URL: http://resources.linkededucation.org/2013/03/lak-dataset-taibi.pdf.