Using Linked Data Traversal to Label Academic Communities

Ilaria Tiddi, Mathieu d'Aquin, Enrico Motta
Knowledge Media Institute, The Open University
Walton Hall, Milton Keynes, MK76AA, United Kingdom
{ilaria.tiddi, mathieu.daquin, enrico.motta}@open.ac.uk

ABSTRACT

In this paper we exploit knowledge from Linked Data to ease the process of analysing scholarly data. In the last years, many techniques have been presented with the aim of analysing such data and revealing new, unrevealed knowledge, generally presented in the form of "patterns". However, the discovered patterns often still require human interpretation to be further exploited, which might be a time and energy consuming process. Our idea is that the knowledge shared within Linked Data can actuality help and ease the process of interpreting these patterns. In practice, we show how research communities obtained through standard network analytics techniques can be made more understandable through exploiting the knowledge contained in Linked Data. To this end, we apply our system Dedalo that, by performing a simple Linked Data traversal, is able to automatically label clusters of words, corresponding to topics of the different communities.

Categories and Subject Descriptors

I.5 [**Pattern Recognition**]; I.5.4 [**Pattern Recognition**]: Application—*Text Analysis*; J.5 [**Computer Application**]: Administrative data processing—*Education*

Keywords

Linked Data, Educational Data, Community Detection

1. INTRODUCTION

Interest for research and scholarly data has sensibly increased in the last years, due to the large and constantly increasing amounts of published data. Many techniques have been applied and presented in the literature in order to mine and visualise such data, with the aim to reveal unrevealed knowledge, highlighting hidden patterns (to be intended, as in [4], as "a statement describing an interesting relationship among a subset of the data") and forecasting interesting trends.

However, the interpretation of the revealed knowledge is still an intensive process, since it requires the intervention of a human expert, whose role is to analyse the trends and give them a meaning before they can be further exploited. This makes interpretation a crucial step in the process, where some knowledge might still remain unrevealed.

The use-case we adopt here is the detection of topic communities within data provided by our university; i.e. a corpus of thousands of papers that have been published by each faculty of the Open University in recent years[1]. For the simple aim of detecting which research areas are being studied, we process documents using basic text-mining techniques to obtain groups of similar documents, corresponding more or less to research areas. The techniques generally employed for purposes like ours tend to probabilistically extract topics as groups of co-occuring words, which eventually need a human to interpret and label them with the right research area.

The nature of Linked Data can facilitate the process of understanding scholarly data: the idea we bring here is not only that educational data are one of the biggest portions within Linked Data (as reported in April 2014[2]), but also that the structured and linked form they are represented with allows the spanning of datasets and the discovery of unrevealed knowledge about them with very little effort. We highlight the Web of (Linked) Data potential of linking RDF datasets across different disciplines, making new sources of knowledge accessible by the machines but also allowing the discovery of unrevealed, multi-domain knowledge. With such an amount of information shared through Linked Data, it should therefore be possible to automatise, or at least facilitate, the interpretation of results such as the ones described above.

What we intend to achieve in this work is automatising the interpretation of topic communities, by using the Linked Data connected information as background knowledge. In this paper, we use an automatic framework traversing Linked Data, Dedalo [13], that uses an A* search strategy over the graph of Linked Data, to identify common explanations (labels) for the the research communities that it found.

2. RELATED WORK

Our work finds its place at the intersection between the *semantic publishing* field, which comprehends Semantic Web-based approaches enriching published data and facilitating

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742019.

[1] http://oro.open.ac.uk/

[2] http://data.dws.informatik.uni-mannheim.de/lodcloud/2014/

their analysis, and the subfield of social network analysis defined as *topical community detection*, consisting in those approaches using the documents' textual information to detect topics of the identified communities.

Semantic publishing. The pioneer works to explore academic data have been built based on citation-indexes. Among those, we include the very well known Google Scholar[3], the DBLP database[4] and the CiteSeer^X [8] search engine. Arguing that those were only focusing on data exploration, a more recent generation including the Microsoft Academic Search[5] and ArnetMiner [12] systems has highlighted the importance trend discovery and prediction, and proposed novel features for those purposes. Recently, the Rexplore [10] system pointed out that the lack of semantic information in the former works prevents a proper data exploration in a granular way, and introduced information from external Linked Data datasets such as GeoNames, DBpedia or DBLP++ to overcome this issue.

The importance of the Semantic Web for scholarly data has also been highlighted in the literature through the use of ontologies and vocabularies to enhance the representation of those data. A whole set of vocabularies, now available as the *SPAR* suite[6], include ontological models in OWL 2.0 DL for publishing and referencing bibliographic records and documents in various aspects of the publication process. Since those models did not take into account the time factor (e.g. author's role changing), the work of [11] presented two ontologies including the *time-indexed value in context* ontology pattern tackling this issue.

Topical community detection. Relevant literature includes a wide range of works for topic labelling (for a full survey on the area, see [3]). Our work is part of those works that make use of external datasources to label topics. In [1, 2, 7], topics are extracted from Wikipedia's structured knowledge, also verifying the topics against a search engine [1, 7]. Moving away from Wikipedia and built-in knowledge bases into Linked Data, [5] proposes to use DBpedia categories as labels for topics. This work is certainly the most similar to our research, but with a significant difference: [5] relies on the use of SPARQL queries to retrieve the DBpedia categories. This introduces some (human) a priori knowledge, and limits the benefits of the Linked Data interconnected knowledge, intended as a more serendipitous knowledge discovery process.

3. PROBLEM STATEMENT

To detect communities that talk about similar things, we can perform clustering on the set of available documents. Given a dataset $\mathcal{D} = \{d_0,\ldots,d_m\}$ of m publications and a corpus $\mathcal{W} = \{w_0,\ldots,w_n\}$ of n words occurring in each $d_k \in \mathcal{D}$, a community is defined as a group of similar words $\mathcal{C} = \{w_0, \ldots, w_j\}$ (where $\mathcal{C} \subseteq \mathcal{W}$) associated to a topic $\mathcal{T}$, that we aim at defining automatically by estimating it on the words' similarity (see next section for details). Once obtained, we can use the 10 words $top_{10}(\mathcal{C}_i) \subseteq \mathcal{C}_i$ that are closest to the centroid of the cluster $\mathcal{C}_i$ to label the community.

Figure 1 shows clusters representing our university's communities. We obtained a network of communities (the size represents the number of documents belonging to it) whose connections intensity shows their relatedness (the stronger is the connection, the thicker is the line). The network reveals indeed different areas; however, quickly identifying communities is hard since, unless being an expert of the domain, words in each cluster remain meaningless. For instance, we distinguish a red group that we call $\mathcal{C}_{y1}$, $\mathcal{C}_{y2}$, $\mathcal{C}_{y3}$, for which:

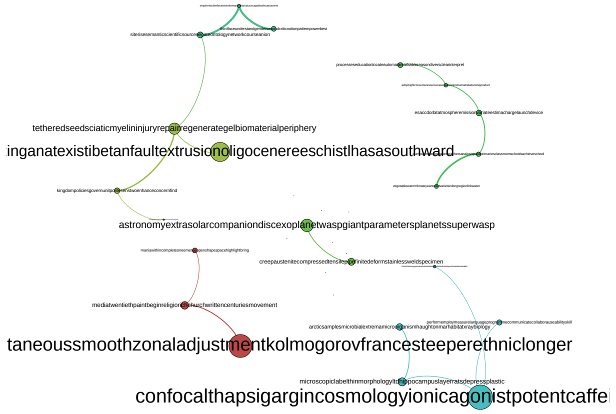
empireresultreflectextentformat...produceecapableafricaexamine
siterisesemanticscientificsourcenegateontologynetworkcourseanion
thirdfaceunderstandgendersecondcriticnotionpatternpowerbest
processeseducationlocateautomatamarkablereasondiversclearinterpret
adoptrightconsumeresourcecase...bookindustrialrelationshipproduct
esaccdorbitatmosphereemissionvariateestimachargelaunchdevice
tetheredseedsciaticmyelininjuryrepairregenerategelbiomaterialperiphery
inganatexistibetanfaultextrusionoligocenereeschistlhasasouthward
primariesclassroomschoolsachieveschool
vegetablewarmclimateyearsa...varietieslongregionlinkwater
kingdompoliciesgovernunitpolitiesenstwoenhanceconcernfind
astronomyextrasolarcompaniondiscexoplanetwaspgiantparametersplanetssuperwasp
maniawithincompleteoneemergeopenshapespacehighlightbring
creepaustenitecompressedtensilepipefinitedeformstainlessweldspecimen
mediatwentiethpaintbeginreligionchchurchwrittencenturiesmovement
arcticsamplesmicrobialextremamicroorganismhaughtonmarhabitatxraybiology
performemploymeasurelanguageprogrammecommunicatecollaborauseabilityskill
taneoussmoothzonaladjustmentkolmogorovfrancesteeperethniclonger
microscopiclabelthinmorphologyltdhippocampuslayerratsdepressplastic
confocalthapsigargincosmologyionicagonistpotentcaffei

Figure 1: The Open University community network.

- $top_{10}(\mathcal{C}_{y1})$ = *media, twentieth, paint, begin, religion, rich, church, write, century, movement*
- $top_{10}(\mathcal{C}_{y2})$ = *inertia, instantaneous, smooth, zonal, adjustment, kolmogorov, france, steeper, ethnic, longer*
- $top_{10}(\mathcal{C}_{y3})$ = *mania, within, complete, one, emerge, open, shape, space, highlight, bring*

To interpret them, one needs to be an expert that, using his own background knowledge, defines a super concept relating them. And even so, this might not be enough to explain the whole community. One could say that $\mathcal{C}_{y1}$ and $\mathcal{C}_{y2}$ might correspond respectively to arts and mathematics, but $\mathcal{C}_{y3}$ would probably remain unexplained to many people.

In [13] we presented Dedalo, a framework to explain clusters of items using knowledge extracted from Linked Data. Dedalo is based on three main assumptions:

- if items are in the same cluster, there is an underlying characteristics that makes items appearing together, and this goes beyond the clustering process;
- Linked Data knowledge is a graph of URI entities connected through RDF properties, that can be blindly navigated in order to serendipitously discover new knowledge (possibly across different datasources), using a simple Linked Data traversal and URI dereferencing process;
- some entities in this graph can have a common walk $\overrightarrow{w}$ (expressed in the form of a chain of contiguous RDF properties) to a specific entity.

Given those assumptions, the main insight is that if items of a cluster share the same walk to a specific (unknown) entity in the Linked Data graph, then these walks can be used as

[3] http://scholar.google.com

[4] http://www.informatik.uni-trier.de/ ley/db/

[5] http://academic.research.microsoft.com/

[6] http://purl.org/spar

an explanation to their grouping. Dedalo then applies an A* graph search strategy, aiming at finding the least-cost path from the set of initial nodes to a goal node, i.e. the entity they have in common somewhere in the graph, and uses the entropy measure to estimate the costs of the walks in the graph. Because Linked Data can be traversed by URI dereferencing, Dedalo explores the graph trying to improve the accuracy of the explanations by iteratively deepening the graph exploration.

Our general challenge is summarised as follows: given the cluster of words, whose understandability remain difficult, Linked Data, providing information about concepts in multivariate domains and Dedalo, which is able to find Linked Data explanations for the grouping of some patterns, we want to set up a process to automatically label communities and ease the process of their interpretation.

4. APPROACH

To label communities, we performed three tasks: (i) data pre-processing, (ii) network building and (iii) community labelling.

4.1 Data pre-processing

The first step consisted in pre-processing the input data. We started from a corpus of publication abstracts $\mathcal{D} = \{d_0,\dots, d_n\}$ and applied common text preprocessing steps to clean them. We intentionally chose the abstracts for their accessibility, as well as because we considered they were enough to represent the research topic of a paper. The use of full texts is left for future work.

(1) Text normalisation. This includes reducing words to lower case as well as removing (English) stopwords, numbers and punctuation.
(2) Stemming and stem completion. Words are first reduced to their stemma, and then each stem is replaced with its shortest possible raw form in $\mathcal{W}$. This improves the words readability and the chances to map them with the same one DBpedia entity. For instance, the words *religion*, *religious* and *religiously* are first all stemmed as *religi-*, and this one is then transformed to *religion*.
(3) Term filtering. We set the minimum characters length for a term w_i to 3, as we considered words below this boundary as pointless to our purposes. This, of course, is a choice purely adapted to our data, and might not be applicable to a different dataset.
(4) DBpedia lookup. We removed from $\mathcal{W}$ words that could not be mapped with DBpedia. Because Dedalo relies on link traversal, we do not have to worry about words ambiguity nor ranking the $top(k)$ relevant DBpedia entities for a word: either a DBpedia entity exists (as in the triple ⟨db:Religion,dc:subject,db:category:Religion⟩), and therefore Dedalo would normally dereference it by collecting its properties and values, or the entity has a redirection property (expressed in DBpedia by the *dbo:wikiPageRedirects* property) that Dedalo would naturally follow as any other property, as when discovering the triple ⟨db:Religiosity,dbo:wikiPageRedirects,db:Religion⟩.

4.2 Network building

We applied the mathematical technique of Latent Semantic Analysis (LSA) to extract and infer relations of expected contextual usage of words in texts [6]. Words have been represented first as high dimensional vectors, so that we obtained a TF-IDF weighted term-document matrix M in which each column was a unique word and each row a document of the corpus. We cropped M at its upper and lower boundaries, i.e. removing words appearing in more than 25% of $\mathcal{D}$, or less than twice. With respect to the power law distribution[7], we considered that boundaries would be helpful in detecting the truly meaningful words with respect to our data.

Secondly, the matrix was reduced into a lower dimensional space, the *latent semantic space*, using a form of factor analysis called the *single value decomposition* (SVD). This space reveals semantic connections between words beyond the lexical level, reproducing the human judgment of meanings similarity. With the SVD, M is first split in three sub-matrices (the term vector *T-matrix*, the document vector *D-matrix* and the diagonal matrix *S-matrix*) and then reduced into a space $\mathcal{S}_k$ of k dimensions, i.e. the latent semantic space. The dimensions reduction collapses the sub-matrices in such a way that words occurring in similar contexts will appear with a greater (or lesser) estimated frequency, therefore automatically reproducing the words' grouping (into what a human would define as a topic).

Once obtained the LSA space S_k, we clustered the words according to the Euclidean distance between them, and formed the set of clusters $\mathcal{C} = \{\mathcal{C}_0, \dots, \mathcal{C}_i\}$ corresponding to the communities. To highlight the communities relatedness, we kept only the connection (the edge) between a given cluster $\mathcal{C}_i$ and its closest one according to the distance between their centroids. The result is a network graph, as already shown in Figure 1, in which clusters are nodes of different size (the number of words belonging to it), and the edges are the $top(1)$ connections between the nodes. The thicker the edge, the closest the two centroids are, the more the two communities are related.

4.3 Communities labelling

The last step is to run Dedalo on each cluster $\mathcal{C}_i$ in order to find explanations revealing why its words w_i are part of it. Dedalo's process is inspired by the Inductive Logic Programming approach [9], in which, starting from a group of positive and negative examples (in our case, words $w_i \in \mathcal{W}$) and some knowledge about them, a set of theories entailing all the positive and none of the negative examples are automatically derived. In this context, given:

- $\mathcal{C}_i$, a cluster of words corresponding to the community that we want to label;
- $\mathcal{W} \backslash \mathcal{C}_i$, the remaining words w_i in $\mathcal{W}$ to use as counter examples;
- $\mathcal{B}$, the knowledge in Linked Data, encoded as walks $\overrightarrow{w}$ of RDF properties between a group of items in $\mathcal{W}$ and a final entity e_i, i.e. $\varepsilon_i = \langle \overrightarrow{w}.e_i \rangle$;

Dedalo's aim is to find the best explanation $top(\varepsilon_i)$ for the words in $\mathcal{C}_i$, where best is intended as representing the biggest number of words of $\mathcal{C}_i$ and the least of the words outside it. An A*-driven link traversal is used to iteratively explore new parts of the Linked Data graph, and to collect the most promising explanations. To start this graph search, we created a URI entity for each word w_i, and linked it

[7] http://en.wikipedia.org/wiki/Zipf_law

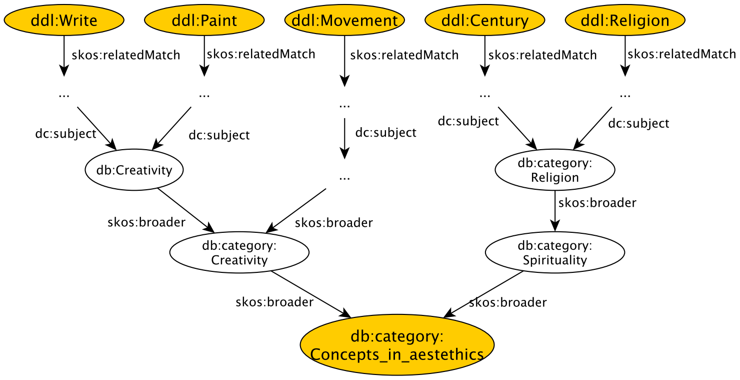

Figure 2: Example of the Linked Data graph traversal. The search starts at the top with the words in $\mathcal{C}_{y1}$. The walks they share in the graph are collected and then evaluated in order to find the most suitable ones.

to its DBpedia correspondent w_i^D(found in the DBpedia lookup step) with a RDF property *skos:relatedMatch*, so that Dedalo's graph initially contains n triples in the form of $\langle$ddl:w_i,skos:relatedMatch,db:$w_i^D\rangle$ with n being the size of the words corpus $\mathcal{W}$.

The graph is iteratively expanded, as follows: first, the best walk $\overrightarrow{w}_i$ is taken from the queue of all the possible walks that could be followed in the graph; second, the entities at the end of this walk are dereferenced; third, new walks of length l+1 (l being the length of the best walk $\overrightarrow{w_i}$) are collected by chaining $\overrightarrow{w_i}$ to the properties obtained by dereferencing the new URIs; finally, those new walks are added to the queue, and a new iteration begins. Each time a new walk $\overrightarrow{w_i}$ is discovered, we also build new explanations ε_i=$\langle\overrightarrow{w_i}.e_i\rangle$, using each of the entities e_i that $\overrightarrow{w_i}$ walks to. Figure 2 gives a non-exhaustive graph search example on $\mathcal{C}_{y1}$. For readability clarity, Table 1 presents a legend of the walks that will be used further on.

Table 1: Walks label of our running example.

id.	$\overrightarrow{w_i}$
$\overrightarrow{w_1}$	{skos:relatedMatch}
$\overrightarrow{w_2}$	{skos:relatedMatch,dc:subject}
$\overrightarrow{w_3}$	{skos:relatedMatch,dc:subject,skos:broader}
$\overrightarrow{w_4}$	{skos:relatedMatch,dc:subject,skos:broader, skos:broader,skos:broader}
$\overrightarrow{w_5}$	{skos:relatedMatch,dc:subject,skos:broader, skos:broader,skos:broader}

If we assume the best walk at a given iteration is $\overrightarrow{w_3}$ ={skos: relatedMatch,dc:subject,skos:broader}, we dereference the entities at the end of $\overrightarrow{w_3}$, i.e. e_1=*db:category:Creativity* and e_2=*db:category:Spirituality*. Thus, we build the new walk $\overrightarrow{w_4}$={skos:relatedMatch,dc:subject,skos:broader,skos:broade r} by adding to $\overrightarrow{w_3}$ the new property p=skos:broader discovered by dereferencing e_1 and e_2, and add it to the queue of walks. Finally, we create a new explanation ε=$\langle\overrightarrow{w_4}$.*db:category:Concepts_in_aestethics*$\rangle$, evaluate it, and start a new iteration.

The explanations accuracy is statistically evaluated using the F-Measure $F = 2 * \frac{P*R}{P+R}$. Given an explanation ε_i = $\langle\overrightarrow{w_i}.e_i\rangle$, Precision and Recall are defined as follows:

$$(1)\ P = \frac{sources(\varepsilon_i) \cap \mathcal{C}_i}{sources(\varepsilon_i)} \qquad (2)\ R = \frac{sources(\varepsilon_i) \cap \mathcal{C}_i}{|\mathcal{C}_i|}$$

where *sources*(ε_i) is the number of words $w_i \in \mathcal{W}$ walking to e_i through the walk $\overrightarrow{w_i}$ and $\mathcal{C}_i$ is the cluster of words we want to explain. For instance, 5 sources are covered by the explanation ε_1=$\langle\overrightarrow{w_4}$.*db:category:Concepts_in_aesthetics*$\rangle$, while the explanation ε_2=$\langle\overrightarrow{w_3}$.*db:category:Creativity*$\rangle$ covers only 3, so we consider ε_1 as the most valuable explanation for the cluster.

Finally, the best explanation is can be used as label of communities, as in Figure 3.

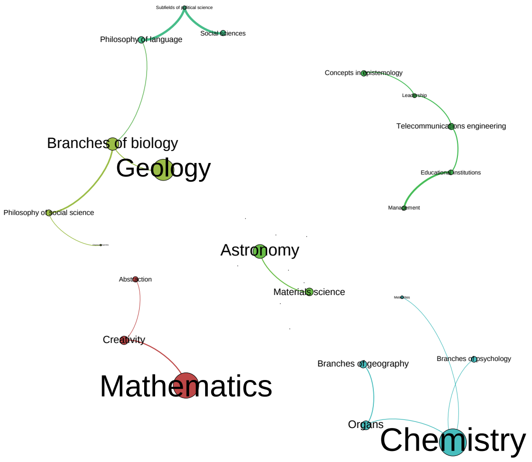

Figure 3: Replacing the clusters of words with the DBpedia categories. Each community is labelled with the best explanation that Dedalo found after 5 iterations.

5. EXPERIMENTS

Below we give some details about our experiments. All the data, experiments and tests are publicly available online[8].

5.1 Data and process details

Data preprocessing. Our dataset $\mathcal{D}$ was composed of 17,142 English abstracts, retrieved from the ORO repository using a simple SPARQL query[9]. The set of words $\mathcal{W}$, initially composed of 65,564 words, was reduced to 18,396 words after the preprocessing step.

[8] http://linkedu.eu/dedalo/

[9] http://data.open.ac.uk/sparql

Table 2: Explanations found by Dedalo after 5 iterations, their F-Measure score (FM), the number of sources $\textit{sources}(\varepsilon_i)$ in $\mathcal{C}_i$ covered by ε_i, and the size of the cluster $\mathcal{C}_i$.

ε_i	FM	$\lvert \textit{sources}(\varepsilon_i) \cap \mathcal{C}_i \rvert$	$\lvert \mathcal{C}_i \rvert$
$\langle \overrightarrow{w_2}.\textit{db:Category:Meteorites}\rangle$	40.0	4	16
$\langle \overrightarrow{w_5}.\textit{db:Category:Geology}\rangle$	32.9	25	125
$\langle \overrightarrow{w_4}.\textit{db:Category:Chemical\_Properties}\rangle$	26.7	2	8
$\langle \overrightarrow{w_5}.\textit{db:Category:Branches\_of\_Biology}\rangle$	26.4	21	73
$\langle \overrightarrow{w_5}.\textit{db:Category:Organs}\rangle$	26.1	9	52
$\langle \overrightarrow{w_4}.\textit{db:Category:Educational\_Institutions}\rangle$	23.5	4	28
$\langle \overrightarrow{w_5}.\textit{db:Category:Chemistry}\rangle$	22.9	28	160
$\langle \overrightarrow{w_5}.\textit{db:Category:Astronomy}\rangle$	22.2	12	81
$\langle \overrightarrow{w_3}.\textit{db:Category:Social\_Sciences}\rangle$	21.3	5	31
$\langle \overrightarrow{w_5}.\textit{db:Category:Mathematics}\rangle$	21.1	22	150
$\langle \overrightarrow{w_5}.\textit{db:Category:Telecommunications\_Engineering}\rangle$	20.8	5	36
$\langle \overrightarrow{w_4}.\textit{db:Category:Subfields\_of\_Political\_Science}\rangle$	20.1	3	21
$\langle \overrightarrow{w_3}.\textit{db:Category:Concepts\_in\_Epistemology}\rangle$	19.5	4	32
$\langle \overrightarrow{w_4}.\textit{db:Category:Creativity}\rangle$	18.2	7	49
$\langle \overrightarrow{w_4}.\textit{db:Category:Abstraction}\rangle$	17.9	6	31
$\langle \overrightarrow{w_4}.\textit{db:Category:Branches\_of\_Psychology}\rangle$	17.9	5	33
$\langle \overrightarrow{w_4}.\textit{db:Category:Management}\rangle$	16.7	3	25
$\langle \overrightarrow{w_5}.\textit{db:Category:Leadership}\rangle$	16.7	3	23
$\langle \overrightarrow{w_5}.\textit{db:Category:Materials\_Science}\rangle$	16.5	7	45
$\langle \overrightarrow{w_4}.\textit{db:Category:Philosophy\_of\_Social\_Science}\rangle$	16.3	4	35
$\langle \overrightarrow{w_5}.\textit{db:Category:Branches\_of\_Geography}\rangle$	14.7	5	42
$\langle \overrightarrow{w_3}.\textit{db:Category:Philosophy\_of\_Language}\rangle$	13.3	3	38

LSA space extraction. To obtain the LSA space $\mathcal{S}_k$ out of $\mathcal{W}$, we used the R LSA package[10]. We produced several k-dimensional spaces $\mathcal{S}$, with k either manually set to 150 and 250 or automatically set to 899. Since no significant difference was seen between the clusters produced by $\mathcal{S}_{150}$, $\mathcal{S}_{250}$ and $\mathcal{S}_{899}$, we chose $\mathcal{S}_{250}$ as a trade-off among them.

Clustering. We used the K-means algorithm and the Weka tool[11] to cluster the words represented in $\mathcal{S}_{250}$ and obtain communities of words. In order to test different granularities, we ran several tests, with the clusters number K set to 20, 30, 50, 100 and 150. For the communities visualisation and as a running example of this work, we chose the $K = 30$ as it was giving a good idea of the connections between communities. Some of Dedalo's explanations with more fine-grained or general clusters are presented in the next section, while the full results are available online. We filtered out of the process clusters whose size $|\mathcal{C}_i|$ was less than 10 or above 500 elements, as we considered them noise. The final $\mathcal{W}$ corpus consisted of 1,192 words.

Networking. The network graph was obtained using the Gephi tool[12].

5.2 Experiments and discussion

Our evaluation is focused on showing the benefits of including Dedalo's strategy to label clusters.

Improvement over iterations. As explained in the previous section, Dedalo iteratively builds a graph and finds explanations while traversing Linked Data. This means that the more the graph is traversed, the more an explanation is likely to be shared by a bigger number of elements, and therefore to improve its accuracy. In Figure 2, for instance, we can see that the $\varepsilon_1 = \langle \overrightarrow{w_3}.\textit{db:category:Creativity}\rangle$ covers three items of $\mathcal{C}_i$, while $\varepsilon_2 = \langle \overrightarrow{w_4}.\textit{db:category:Concepts\_in\_aesthetics}\rangle$ covers 5 of them. Table 3 gives an overview of the explanations improvement for the clusters $\mathcal{C}_{y1}$, $\mathcal{C}_{y2}$ and $\mathcal{C}_{y3}$, by showing the best explanation at each iteration, as well as its F-Measure.

As one can see, within few iterations we automatically span from our initial dataset to DBpedia, and manage to build explanations that generalise the clusters and explain why words appear together. Dedalo's A* search detects in a first instance that the DBpedia property *dc:subject* is the most promising walk (where promising means the one that is more likely to reveal a good explanation in terms of F-Measure), followed by the property *skos:broader*. For this reason, most of the explanations after few iterations have already walked up the taxonomy of DBpedia concepts by following two or three *skos:broader* properties (shown by the walks' apex in the Table). With this strategy, we can see how the explanation for a cluster significantly improves in a short time (we pass from "*2% of the words in* $\mathcal{C}_{y2}$ *match the DBpedia concept* db:Scale" to "*20% of the words in* $\mathcal{C}_{y2}$ *are subcategories of the category* Mathematics").

Table 3: Example of explanations for $\mathcal{C}_{y1}$, $\mathcal{C}_{y2}$ and $\mathcal{C}_{y3}$ found at each iteration.

$\mathcal{C}_i$	iter	best explanation ε	F(%)
$\mathcal{C}_{y1}$	1	$\langle \overrightarrow{w_1}.\textit{db:Century}\rangle$	7.8
	2	$\langle \overrightarrow{w_2}.\textit{db:Category:Writing}\rangle$	11.3
	3	$\langle \overrightarrow{w_3}.\textit{db:Category:Problem\_solving}\rangle$	13.5
	4	$\langle \overrightarrow{w_4}.\textit{db:Category:Creativity}\rangle$	18.2
$\mathcal{C}_{y2}$	1	$\langle \overrightarrow{w_1}.\textit{db:Scale}\rangle$	2.6
	2	$\langle \overrightarrow{w_2}.\textit{db:Category:Concepts\_in\_Physics}\rangle$	10.3
	3	$\langle \overrightarrow{w_3}.\textit{db:Category:Physics}\rangle$	14.3
	4	$\langle \overrightarrow{w_4}.\textit{db:Category:Fields\_of\_Mathematics}\rangle$	18.5
	5	$\langle \overrightarrow{w_5}.\textit{db:Category:Mathematics}\rangle$	21.1
$\mathcal{C}_{y3}$	1	$\langle \overrightarrow{w_1}.\textit{db:Social}\rangle$	6.0
	2	$\langle \overrightarrow{w_2}.\textit{db:Category:Concepts\_in\_Logics}\rangle$	10.8
	3	$\langle \overrightarrow{w_3}.\textit{db:Category:Logic}\rangle$	15.0
	4	$\langle \overrightarrow{w_4}.\textit{db:Category:Abstraction}\rangle$	17.9

[10] http://cran.r-project.org/web/packages/lsa/lsa.pdf

[11] http://www.cs.waikato.ac.nz/ml/index.html

[12] https://gephi.github.io/

Fine-grained clusters labelling. Table 2 shows the best explanation that has been found for each of the 23 clusters after 5 iterations. The others columns are: F-Measure, the numbers of sources covered by the explanations and the size of the cluster $\mathcal{C}_i$. Dedalo exploits Linked Data knowledge to give an automatic label to each community, and we can see that labels do not only give more sense to the groups of words, but also reflect the distinction of different research areas. Those labels facilitate the user's analysis: for instance, labelling the three clusters $\mathcal{C}_{y1}$, $\mathcal{C}_{y2}$ and $\mathcal{C}_{y3}$ respectively as *Creativity*, *Mathematics* and *Abstraction* reveals an hidden connection between the communities that could not be that visible simply by using the cluster's top_{10} words.

Finally, we can observe how the labelling process reflects the granularity of the clustering process: the more clusters we create, the more fine-grained is the explanation; while the less there are, the more general the topic is. For instance, for $K = 20$, we observed a community whose best explanation is: $\varepsilon_1 = \langle \overrightarrow{w_5}.\textit{db:Category:Culture} \rangle$ (13% F-Measure), while we notice that its second and third best ones, $\varepsilon_2 = \langle \overrightarrow{w_3}.\textit{db:Category:Philosophy\_of\_Language} \rangle$ and $\varepsilon_3 = \langle \overrightarrow{w_4}.\textit{db:Category:Social\_Science} \rangle$ do correspond to the labels of two different clusters when $K = 30$. Inversely, for $K = 50$, we have obtained different communities, each of one explained by mathematics subcategories, for instance, (i) $\varepsilon_1 = \langle \overrightarrow{w_4}.\textit{db:Probability\_and\_statistics} \rangle$, (ii) $\varepsilon_2 = \langle \overrightarrow{w_2}.\textit{db:Category:Elementary\_Mathematics} \rangle$, and (iii) $\varepsilon_3 = \langle \overrightarrow{w_4}.\textit{db:Category:Reasoning} \rangle$.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we presented a use-case for labelling scholarly data represented as groups of related words, by exploiting knowledge from Linked Data. To achieve this, we used Dedalo, a system able to find explanations from Linked Data for a group of items using a graph search strategy and a Linked Data traversal. We automatically obtained labels for a group of "semantically related words" that otherwise would have required the experts background knowledge to be explained. We have shown that the explanations serendipitously found by Dedalo can ease the process of understanding the words grouping. The result is a more human-readable network of academic communities that only relies on the knowledge contained in Linked Data.

As future work, we will consider other approaches to exploit Dedalo and Linked Data to ease the analysis of published data, as well as using them for prediction purposes. Another axis of research is combining the explanations to obtain a more precise explanation of the cluster.

7. ACKNOWLEDGEMENT

Authors would like to thank E. Bastianelli (Semantic Analytics Group at Uniroma2) for the fruitful discussion and help in outlining the early stage of this work.

8. REFERENCES

[1] N. Aletras and M. Stevenson. Labelling topics using unsupervised graph-based methods.

[2] K. Coursey, R. Mihalcea, and W. Moen. Using encyclopedic knowledge for automatic topic identification. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 210–218. Association for Computational Linguistics, 2009.

[3] Y. Ding. Community detection: Topological vs. topical. *Journal of Informetrics*, 5(4):498 – 514, 2011.

[4] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37, 1996.

[5] I. Hulpus, C. Hayes, M. Karnstedt, and D. Greene. Unsupervised graph-based topic labelling using dbpedia. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 465–474. ACM, 2013.

[6] T. K. Landauer, P. W. Foltz, and D. Laham. An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284, 1998.

[7] J. H. Lau, K. Grieser, D. Newman, and T. Baldwin. Automatic labelling of topic models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1536–1545. Association for Computational Linguistics, 2011.

[8] H. Li, I. Councill, W.-C. Lee, and C. L. Giles. Citeseerx: an architecture and web service design for an academic document search engine. In *Proceedings of the 15th international conference on World Wide Web*, pages 883–884. ACM, 2006.

[9] S. Muggleton. *Inductive logic programming*, volume 168. Springer, 1992.

[10] F. Osborne, E. Motta, and P. Mulholland. Exploring scholarly data with rexplore. In *The Semantic Web–ISWC 2013*, pages 460–477. Springer, 2013.

[11] S. Peroni, D. Shotton, and F. Vitali. Scholarly publishing and linked data: describing roles, statuses, temporal and contextual extents. In *Proceedings of the 8th International Conference on Semantic Systems*, pages 9–16. ACM, 2012.

[12] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: Extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 990–998, New York, NY, USA, 2008. ACM.

[13] I. Tiddi, M. d'Aquin, and E. Motta. Dedalo: Looking for clusters explanations in a labyrinth of linked data. In *The Semantic Web: Trends and Challenges*, pages 333–348. Springer, 2014.