Unsupervised Graph-based Topic Labelling using DBpedia

Ioana Hulpuş, Conor Hayes,
Marcel Karnstedt
Digital Enterprise Research Institute (DERI)
National University of Ireland
Galway, Ireland
{*first.last*}@deri.org

Derek Greene
School of Computer Science and Informatics
University College Dublin
Dublin, Ireland
Derek.Greene@ucd.ie

ABSTRACT

Automated topic labelling brings benefits for users aiming at analysing and understanding document collections, as well as for search engines targetting at the linkage between groups of words and their inherent topics. Current approaches to achieve this suffer in quality, but we argue their performances might be improved by setting the focus on the structure in the data. Building upon research for concept disambiguation and linking to DBpedia, we are taking a novel approach to topic labelling by making use of structured data exposed by DBpedia. We start from the hypothesis that words co-occuring in text likely refer to concepts that belong closely together in the DBpedia graph. Using graph centrality measures, we show that we are able to identify the concepts that best represent the topics. We comparatively evaluate our graph-based approach and the standard text-based approach, on topics extracted from three corpora, based on results gathered in a crowd-sourcing experiment. Our research shows that graph-based analysis of DBpedia can achieve better results for topic labelling in terms of both precision and topic coverage.

1. INTRODUCTION

One of the most popular approaches for identifying the subject matter of a collection of documents is to determine its inherently addressed topics. Several methods have been proposed for probabilistic topic modelling, such as Latent Dirichlet Allocation (LDA) [2], Pachinko Allocation [11] or Probabilistic Latent Semantic Analysis (pLSA) [7]. They model the documents as a mixture of topics, where each topic is treated as a probability distribution over words. As such, topics consist of groups of co-occurring words, ranked by their relevance. Such models are largely used in the domain of text analysis for summarising big document corpora.

Typically users have then to interpret these sets of words in order to label the underlying concepts for further processing and classification. Labelling in this context refers to finding one or a few single phrases, or better concepts, that sufficiently describe the topic in question. This can become a cumbersome task when a corpus is summarised by some hundreds of topics. In this light, automatic topic labelling becomes an important problem to solve in order to support users in their task to efficiently and conveniently analyse, understand and explore document collections. Besides that, it further promises benefits for web search engines, as it allows clustering groups of words under the same umbrella term.

Furthermore, there is an increased interest in research on linking text documents to external knowledge bases that are often created collaboratively by communities and validated by multiple experts. Many benefits are expected to result from this integration, in areas like information retrieval, classification and knowledge discovery and visualisation. One of best known multidomain knowledge bases is DBpedia[1], which extracts structured information from Wikipedia in the form of an openly accessible, consensus driven semantic graph of concepts and relations. This paper describes an approach to automatically extract topic labels by linking the inherent topics of a text to concepts found in DBpedia and mining the resulting semantic topic graphs. Our aim is not only to find a good label itself, but also to integrate the topic into a knowledge base to support subsequent exploitation and navigation of related concepts. An important aspect of our work is therefore to relate a topic label with a URI identifying a concept, which opens the way for facilitating knowledge exploration in DBpedia – and far beyond, based on its rich linkage within Linked Open Data project.

We argue that current approaches for topic labelling based on content analysis capture the essence of a topic only to a limited extent. Mainly, because they do not focus on the structure behind the concepts, nor on the navigation and exploration of these topics. We hypothesise that concepts co-occurring in the text are also closely related in the DBpedia graph. Using graph centrality measures, we are able to identify the concepts that are most likely to represent the topics and are therefore suited to label them. Our contribution can be summarised as follows:

1. We propose a novel approach for topic labelling that relies only on structured data – and provides means to fully exploit its potential. The method does not require any pre-processing and can thus be run directly on-line against queryable knowledge bases like DBpedia.

2. The approach is suited for finding a good label *and* for integrating the topic into a knowledge base to sup-

[1] `http://dbpedia.org`

port subsequent exploitation and navigation of related concepts.

3. We show that graph-based algorithms can be used with success to label topics and that they provide richer knowledge than purely text-based methods.

4. We present a thorough comparative evaluation, based on human judgements about the quality of labels, collected through a crowd-sourcing experiment.

Section 2 discusses related work, and in Section 3 we briefly overview the overall framework that comprises the topic modelling proposed in this work. Based on a motivating example, we formalise the problem statement and introduce the general principle of our solution. We present our approach for graph-based topic labelling in Section 4. We examine particular important aspects of our approach and compare it to the standard text-based approach in terms of precision and topic coverage in Section 5. This evaluation is based on the results from a crowd-sourcing experiment involving texts from three different document corpora. Finally, we conclude in Section 6.

2. RELATED WORK

Several works [15, 12, 10] consider topic labelling in the same scenario as we do, where topics represented by a set of words have to be labelled. A second relevant area considers labelling of document clusters [24, 3, 17, 21]. Similar to the first scenario, document clusters are often summarised as a collection of the most prominent words they contain. The third related direction deals with annotations for indexing [4, 23, 14], also called automatic topic identification [4]. In contrast to identifying exactly one or a couple of labels, this aims for identifying as many as possible concepts that are strongly related to the document in question. Despite these different application domains, the various approaches are better distinguished by the techniques they use.

A significant part of the approaches *extracts* the most likely label from the text, such as [17, 24, 15]. An important drawback is that they rely on the assumptions that (i) the correct label can be found in the documents, and that (ii) the corpus is rich enough to identify a label with confidence. However, this is not always the case. For example, a cluster of documents might be about *artificial intelligence* without mentioning the phrase. On the other hand, it might contain many more specialised phrases that cannot be related just based on the text (e.g., *probabilistic reasoning* and *first-order logic*). This problem can be overcome by the use of external data sources. Besides the work in hand, this idea motivates a wide range of recent research [21, 3, 12, 10, 4, 23, 14].

The probably most popular external knowledge base for this purpose is Wikipedia. Usually, a Wikipedia dump is pre-processed into a local data structure that is subsequently analysed in order to extract suitable labels. [21] manipulates the Wikipedia dump by deconstructing it into a collection of minipages corresponding to subsections of Wikipedia article. The label for the document cluster is selected out of the sections' headings. Besides using Wikipedia, the authors of [10] also query the Google web search engine to obtain label candidates. [4] uses the entire English Wikipedia to build a so-called *encyclopedic graph*. In this graph, the nodes represent Wikipedia articles and their categories. Afterwards, a biased PageRank algorithm is used to weight the nodes of the graph with respect to the queried key-phrases. Unfortunately, the work does not provide enough details to be able to reconstruct the graph. The approach proposed by [23] uses a simplified spreading activation algorithm on the graph consisting of Wikipedia articles and their categories. It relies on the cosine similarity between article texts and the texts of the target documents. While also being graph-based, the work presents only very small scale "informal evaluation". All the aforementioned approaches using Wikipedia strongly differ from our approach, as they analyse the content of Wikipedia articles in order to decide on the proper label. This makes the algorithms hard to adapt to data sources that are less rich in content and do not contain encyclopedic text about concepts Our approach is fully structured and independent of the content of Wikipedia articles.

Another topic-labelling approach using an external data source is [12]. This approach differs from our work and the aforementioned ones by relying on a tree-structured external data source, the Open Directory Project[2]. The authors model each node of the hierarchy as a list of words and compare the topics to label with the nodes in the hierarchy based on various similarity measures. Using a novel algorithm, they select and reuse the label of the most similar node. However, this approach is particularly suited for use-cases providing a given hierarchy that has to match the clusters of the corpus. For less constrained scenarios, as we illustrate in Section 3.2, we see strict usage of tree-shaped knowledge bases as problematic.

Our approach differs from all the above works from three perspectives: First, it uses only structured data in order to identify the labels, which strictly correspond to concepts found in DBpedia. Second, the analysed graphs are not pre-processed off-line. Thus, it can be used entirely on-line by querying knowledge bases, such as the DBpedia SPARQL endpoint[3]. Third, for identifying suitable labels, we adapt and experiment with popular graph-based centrality measures that have not been used before for this task.

3. OVERVIEW

The topic labelling approach proposed in this work is part of a larger framework supporting automated topic analysis, called *Canopy*. Figure 1 illustrates its overall architecture. In this section, we present an overview of the system and formally define the problem that this work focuses on. We also introduce the main notation and terms used throughout the paper.

3.1 The Canopy Framework

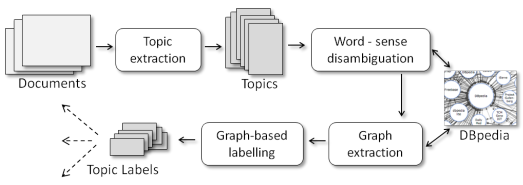


Figure 1: *Canopy* framework for automated topic analysis

[2]http://www.dmoz.org
[3]dbpedia.sparql.endpoint

At the time of writing this paper, Canopy consists of four main components: the topic extraction, the word-sense disambiguation (WSD), the graph extraction and the graph-based labelling. These are the basic building blocks on top of which we envision a range of possible application scenarios: corpus summarisation, visualisation, browsing and extraction of new knowledge by uncovering latent dependencies and relations between topics or domains, etc.

The topic extraction applies the LDA probabilistic topic model [2] to extract *topics* from a corpus of documents. For the sake of generality, we consider each topic θ as a set of *words* (rather than a probability distribution over words). In previous work [9], we have demonstrated a novel graph-based approach to WSD that addresses the linking problem. The WSD determines a set $\mathbf{C}^\theta$ of DBpedia *concepts*, where each $C \in \mathbf{C}^\theta$ represents the identified *sense* of one of the top-k words of a topic θ. Usually, it is neither feasible nor required to relate all top-k words to a concept.

In this paper, we propose a solution for the third and fourth stage of the Canopy process, which together provide the actual topic labelling. In the following, we provide a motivating example and then formalise the underlying problem.

3.2 Example

An intuitive approach for labelling topics represented by a set of concepts $\mathbf{C}^\theta$ is to determine a minimum spanning tree encompassing all $C \in \mathbf{C}^\theta$ from an according hierarchical knowledge base. The least common subsumer in such a tree could then be chosen as a label. However, this would most often produce very generic terms, very close to the root of the overall hierarchy of concepts. This is mainly due to the nature of probabilistic topic models, which do not necessarily group concepts of the same type. Consider a topic θ described by [`patient, drug, hospital, health, professional ...`]. All top five words come from very different branches of a standard knowledge base like WordNet. In this case, the least common subsumer is the very root of the WordNet hierarchy: *Entity*. Similarly, in DBpedia's structure of categories, the least common subsumer is *Life*. However, a considerably good label would be *Health* itself, *Healthcare* or even *Medicine*. *Healthcare* is the least common subsumer of `patient, drug` and `hospital`, and a child of *Health*. *Medicine*, however, is only subsuming `drug`. In order to identify good labels we can thus not rely on the simple least common subsumer. This motivates us to to exploit graph specific methods on graph-based knowledge repositories, in our case DBpedia. Further, it illustrates the main challenges of this approach, which we formalise next.

3.3 Problem Statement and General Approach

In this paper, we consider the task of topic labelling independent of the way the topics have been linked and disambiguated to DBpedia concepts. We formulate the problem as follows: Let $\mathbf{C}^\theta$ be a set of n DBpedia concepts C_i, $i = 1, .., n$, that correspond to a subset of the top-k words representing one topic θ. The problem is to identify the concept $\mathcal{C}^*$ from all available concepts in DBpedia, such that the relation $r(\mathbf{C}^\theta, \mathcal{C}^*)$ is optimised. Thus, the main challenges are:

1. to extract an appropriate set of concepts from DBpedia as candidates for $\mathcal{C}^*$, and
2. to define r, which quantifies the strength of the relation between the concepts $C_i \in \mathbf{C}^\theta$ and $\mathcal{C}^*$, in a way resulting in topic labels that are meaningful for humans.

We propose to extract a good candidate set by extracting a *topic graph* $\mathbf{G}$ from DBpedia consisting of the close neighbours of concepts C_i and the links between them (*graph extraction*). Then, we investigate how to define the relation r by analysing the conceptual graph of DBpedia underlying $\mathbf{G}$. We adopt principles from social network analysis to identify in $\mathbf{G}$ the most prominent concepts for labelling a topic θ (*graph-based labelling*).

4. GRAPH-BASED TOPIC LABELLING

The intuition behind our approach is that as the concepts of a topic are related, they should lie close in the DBpedia graph. This implies that by expanding from each such concept for a few hops, all the topic concepts will ideally form one connected graph. The graph extraction phase uses this inutition to address the problem of finding label candidates. The sense graph of each concept is defined as follows:

Definition 1. The *sense graph* of a concept C_i is an undirected graph $G_i = (V_i, E_i, C_i)$, where V_i is the set of nodes and E_i is the set of edges connecting the nodes. $C_i \in V_i$ is a DBpedia concept called the *seed concept* or *seed node* of graph G_i.

The topic graph $\mathbf{G}$ is a union of the sense graphs of one topic. It is passed as the result of the graph extraction phase to the actual graph-based labelling step.

Definition 2. Let $\mathbf{C}^\theta = \{C_1, .., C_n\}$ be the set of DBpedia concepts corresponding to the disambiguated senses of the words in the latent topic θ, and let $G_i = (V_i, E_i, C_i)$ be the sense graph corresponding to C_i, $\forall i \in 1, .., n$. Then, if $\mathbf{V} = \bigcup V_i$, and $\mathbf{E} = \bigcup E_i$, then $\mathbf{G} = (\mathbf{V}, \mathbf{E}, \mathbf{C}^\theta)$ is called the *topic graph* of θ, and the concepts $C_i \in \mathbf{C}^\theta$ are the *seed concepts* or *seed nodes* of $\mathbf{G}$.

The second problem, to define a measure for assessing the goodness of all candidate topics, is solved by applying adapted graph centrality measures on the topic graph. This is based on the assumption that a good topic label should be a central node in the topic graph, with respect to the seed nodes. We discuss the benefits of different centrality measures, the reasons for adapting and the resulting formulae for computing them in Section 4.3.

4.1 DBpedia Graph Extraction

The objective of the graph extraction is to identify candidate *concepts* from DBpedia suited to label the topic and to provide an entry point for further knowledge exploration. Starting from the seed node, we want to follow edges of certain type (i.e., "properties") to reach other nodes representing candidate concepts (i.e., "entities" identified by URIs). Because the topic labelling can be seen as assigning classes to topics, we focus mainly on the DBpedia class structure. DBpedia provides three different classification schemata for things, which we describe below. While there is overlap between them, each is derived from a different part of Wikipedia using different processes. As such, our approach will combine the data from these schemata to build topic graphs rather than strict concept hierarchies.

Wikipedia Categories The Wikipedia categorisation system provides a valuable source for categorising the concepts found in DBpedia. It contains 740,000 categories whose structure and relationships are represented by the Simple Knowledge Organization System Vocabulary (prefixed by `skos:`) [25]. The linkage between DBpedia concepts and Wikipedia categories is defined using the `subject` property of the DCIM terms vocabulary (prefixed by `dcterms:`) [8]. We can then extract a category's parent and child categories by querying for the properties `skos:broader` and `skos:broaderOf`. This structure is not a proper hierarchy as it contains cycles [1].

YAGO The YAGO vocabulary represents an ontology automatically extracted from Wikipedia and WordNet [6]. It is linked to DBpedia and contains 365,372 classes. Classes are organised hierarchically and can be navigated using the `rdfs:type` property and `rdfs:subClassOf` property. For example, the DBpedia entity *dbres:Elvis_Presley* has property `rdfs:type` *yago:wikicategory_American_rock_singers*, which in turn has a `rdfs:subClassOf` property of *yago:wordnet_singer_110599806*.

DBpedia Ontology The DBpedia ontology is a shallow, cross-domain ontology that has been manually created based on the most commonly used infoboxes in Wikipedia [1]. It contains 320 classes organised into a subsumption hierarchy. In a similar way to the YAGO approach, DBpedia concepts can be navigated following the `rdfs:subClassOf` and `rdfs:type` properties.

Given a topic θ, for each concept $C_i \in \mathbf{C}^\theta$, we extract a sense graph G_i by querying for all nodes lying at most two hops away from C_i, recursively taking into account all existing edges of type `skos:broader`, `skos:broaderOf`, `rdfs:subClassOf`, `rdfs:type` and `dcterms:subject`. We then merge the sense graphs together, obtaining the topic graph $\mathbf{G}$. The decision to use a distance of two-hops was made after several experiments with node expansions. Expanding the nodes of coherent topics to three hops tended to produce very large graphs and introduce a lot of noise.

Figure 2 exemplifies the graph extraction phase. At the top, it shows four sense graphs for a topic consisting of four DBpedia concepts or 'resources' (prefixed as `dbres:`): `dbres:Energy`, `dbres:Atom`, `dbres:Electron` and `dbres:Quantum`. The dark nodes represent the seed nodes corresponding to these concepts. At the bottom, the figure shows the topic graph obtained by merging them.

One problem we face in the DBpedia graph is that concepts are often linked with Wikipedia administrative categories (e.g., `Category:Pages_containing_deleted_templates`), nodes referring to ethimology (e.g., `Category:Latin_loanwords`) and with very generic LOD concepts (e.g., `owl:Thing`, `owl:Class`, `skos:core#Concept`, etc.). These nodes create a range of shortcuts between concepts that do not reflect relationships we are interested in. For example, if all concepts are considered an instance of `skos:core#Concept`, then there will be a path of length two between *any* two concepts.

To overcome this, we automatically created a list of *stop URIs* that tries to cover this type of nodes. We created that list by navigating the higher levels of the category hierarchy rooted at the node `Category:Contents`. We made the list of 865 identified stop URIs publicly available[4], as it was generated in May 2012.

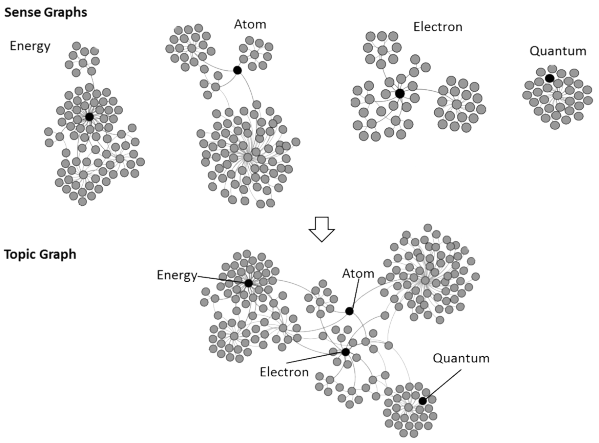


Figure 2: Four sense graphs merged into one topic graph.

4.2 Graph Connectivity

An important hypothesis in our work is that the sense graphs of one topic are more likely to become connected than random concepts in DBpedia. In order to validate this, we ran some experiments on the ground truth data we collected in a previous user study on concept linking and disambiguation [9]. This data consists of 111 topics that had all the top-7 words manually linked and disambiguated by human users against DBpedia concepts or WordNet synsets. We used only the DBpedia concepts and computed for each topic a measure of pairwise concept seed connectivity after two hops.

$$PairConnectivity^{\mathbf{C}^\theta} = \frac{\sum_{C_i \in \mathbf{C}^\theta; C_j \in \mathbf{C}^\theta} \mathbf{1}\{V_i \cap V_j \neq \emptyset\}}{|\mathbf{C}^\theta|(|\mathbf{C}^\theta| - 1)}$$

where $\mathbf{1}\{.\}$ represents the indicator function, V_i/V_j represent the set of nodes of the sense graphs seeded by C_i/C_j.

We compared the obtained average Pair Connectivity over all 111 topics to the same measure applied to identical 111 groups of DBpedia concepts formed by randomising the initial topics, and inserting random noise. For the case of ground truth concepts we obtained an average of 0.46 with standard deviation 0.09, while the random groups had an average pair connectivity of 0.07 and standard deviation 0.02. These values indicate that the connectivity of seed concepts obtained from correctly linking and disambiguating topics to DBpedia is not accidental. Throughout the remaining of this paper, we consider seed concepts that are not connected to the main component as noise, and ignore them from our computation. At the same time, we introduce the term *core concept* to refer to the seed concepts that belong to the main connected component, which similarly is called *core component*.

Figure 3 gives an example on the evolution of the topic words towards the topic labels. In the illustrated example, the topic consists of fifteen words. Out of these words, only eleven *seed concepts* were obtained after linking and disambiguating to DBpedia. Further on, after the graph extraction that we just described, only the nine underlined concepts became connected in the graph, so they became

[4] `http://uimr.deri.ie/sites/StopUris`

core concepts. The other two seed concepts remained disconnected. This exemplifies how noise concepts (i.e. resulting from wrong disambiguation) are implicitly isolated.

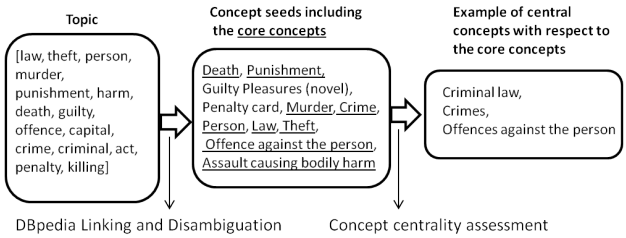


Figure 3: Evolution from topic words to candidate labels.

Intuitively, nodes central to the core component offer themselves as good labels for the topic. In the following section, we focus on defining this "centrality" and detailing our suggested approach for identifying the most central concepts with respect to the core concepts.

4.3 Centrality Measures and Labelling

As previously discussed, relying upon subsumption relationships tends to produce very generic topic labels. Instead our approach is to build a topic graph from which we can analyse the contributions that each node makes to the structure of the graph. As our graph is a semantic graph, we hypothesise that nodes that play important structural role in the graph also have an important semantic relationship to the seed concepts. We select our candidate labels from these nodes. In this section, we discuss how several centrality measures proposed in literature are suited for this task and introduce adaptations required in our case. We use the example in Table 1 throughout this section to illustrate our argumentation. It shows the top-3 concepts that the methods discussed below produce for one example topic.

Centrality measures are a well-known concept in (social) network science. They are used to identify nodes (or actors) that are most important (and thus, central) for the network, an objective in line with our own requirements. Different criteria for importance, suitable for different purposes and scenarios, led to a range of centrality measures proposed in the literature [20]. Two of the most popular ones are:

Closeness centrality: a node is important if it lies close to all of the other nodes in the network. In the context of topics, nodes with high closeness centrality indicate concepts that are closely related to all other concepts of the topic graph.

Betweenness centrality: a node is important if it facilitates the flow of information between other nodes in the graph. In a semantic network, nodes with high betweenness centrality are the nodes that establish short connections between the other nodes in the graph.

These properties intuitively recommend themselves for identifying labels. However, particularly the betweennness centrality is strongly biased towards nodes with high degree, or nodes that are central in large local groups of nodes. This holds even stronger for another centrality measure, the degree centrality, which directly reflects the node degrees. Table 1 shows that "Chemistry" is ranked high by both measures. This concept, part of `Atom`'s sense graph, lies at the center of the large node group in the top right of the topic graph in Figure 2.

In general, centrality measures compute the importance of a node with respect to all other nodes. This means, all network nodes contribute with the same weight to its score. However, in the case of topic labelling, we are particularly interested in the seed concepts, as it is their combination that defines the topic. We therefore propose to adapt the centrality measures so that they focus on the seed nodes, rather than to all the nodes of the graph. We call these adapted measures *focused centralities*. This focus on seed nodes reduces the impact that broad concept nodes have due to their high degree and that of dense local clusters due to their sheer size – as explained above for Figure 2. Table 1 illustrates that the focused variants indeed determine, in comparison to their non-focused counterparts, concepts that are more related to the seed concepts.

Although popular, closeness and betweenness centrality rely on shortest paths only. The assumption that the spread of information is best modelled by the use of shortest paths has been questioned [19, 22]. Some alternatives have been suggested, which consider all paths in a network rather than just the shortest paths. Two such measures are:

Information centrality [22]: Related to the closeness centrality, the information of a path is inverse proportional to its length. This measure aggregates the information of all paths connecting a node with other nodes.

Random walk betweenness centrality [19]: As its name suggests, this measure is a variation of the betweenness centrality. It roughly measures how often a node is traversed by a random walker going from any node in the network to another.

Again, Table 1 illustrates the impact these measures can have. Information centrality and random walk betweenness centrality rank *Particle Physics* highest. This is a more discriminative and therefore better candidate for labelling the example topic than the rather broad *Fundamental Physics Concepts*, ranked highest by the variants based on only shortest paths.

In the following, we explain how to compute the finally chosen set of centrality measures. For simplicity, we assume that the topic graph $\mathbf{G}$ consists of the core component only.

Focused Closeness Centrality: fCC. The average shortest distance l_i from a node i to all other nodes in a graph $G = (V, E)$ is computed as

$$l_i = \frac{1}{|V|} \sum_{v_j \in V} d_{ij} \quad ,$$

where d_{ij} is the length of the shortest path between nodes v_i and v_j. The closeness centrality CC is calculated as the inverse of this average:

$$CC_i = \frac{1}{l_i} = \frac{|V|}{\sum\limits_{v_j \in V} d_{ij}} \quad .$$

In our case, the adapted *focused closeness centrality* fCC is computed as:

$$fCC'_i = \begin{cases} \frac{n}{\sum\limits_{C_j \in \mathbf{C}^\theta} d_{ij}} & v_i \notin \mathbf{C}^\theta; \\ \frac{n-1}{\sum\limits_{C_j \in \mathbf{C}^\theta} d_{ij}} & v_i \in \mathbf{C}^\theta; \end{cases} \quad ,$$

where n is the number of seed nodes in $\mathbf{G}$. Note that if $v_i \in \mathbf{C}^\theta$, there are only $n-1$ other seed nodes in $\mathbf{C}^\theta$ and thus we use $n-1$ as denominator.

Rank	Degree	Closeness Centrality on all graph (CC)	Focused Closeness Centrality (fCC)	Focused Information Centrality (fIC)	Betweenness Centrality on all graph (BC)	Focused Betweenness Centrality (fBC)	Focused Random Walk Betweenness (fRWB)
1	Chemistry	Thermodynamic Properties	Fundamental Physics Concepts	Particle Physics	Chemistry	Fundamental Physics Concepts	Particle Physics
2	Energy	Thermodynamic Free Energy	Physics	Fundamental Physics Concepts	Fundamental Physics Concepts	Particle Physics	Quantum Mechanics
3	Quantum Mechanics	Orbits	Classical Mechanics	Quantum Mechanics	Energy	Quantum Mechanics	Fundamental Physics Concepts

Table 1: Example top-3 labels for topic: [`atom, energy, electron, quantum, classic, orbit, particle`]

Focused Information Centrality: fIC. The information centrality in a graph $G = (V, E)$ is computed as follows:

1. Define a $|V| \times |V|$ matrix B containing the elements:

$$b_{ij} = \begin{cases} 0 & \text{if } v_i \text{ and } v_j \text{ are incident} \\ 1 & \text{otherwise} \end{cases}$$
$$b_{ii} = 1 + degree(v_i) \quad .$$

2. The information contained in the combined path between v_i and v_j is given by:

$$I_{ij} = (c_{ii} + c_{jj} - 2c_{ij})^{-1} \quad ,$$

where c_{ij} are the elements of the matrix $C = B^{-1}$.

3. For a node v_i the information centrality IC is then computed as:

$$IC_i = \frac{|V|}{\sum_{v_j \in V} 1/I_{ij}} \quad .$$

In our case, the *focused information centrality* fIC is computed as:

$$fIC_i = \begin{cases} \frac{n}{\sum_{C_j \in \mathbf{C}^\theta} 1/I_{ij}} & v_i \notin \mathbf{C}^\theta; \\ \frac{n-1}{\sum_{C_j \in \mathbf{C}^\theta} 1/I_{ij}} & v_i \in \mathbf{C}^\theta; \end{cases} \quad .$$

Focused Betweenness Centrality: fBC. For the betweenness centrality, we have to be aware that for every pair of nodes there might exist several shortest paths that pass through the node of interest. The betweenness centrality BC of a node v_i in a graph $G = (V, E)$ is computed as:

$$BC_i = \frac{\sum_{v_s, v_t \in V \wedge s<t} \frac{x^i_{st}}{g_{st}}}{|V|(|V|-1)/2} \quad ,$$

where x^i_{st} is the number of shortest paths between v_s and v_t that pass through node v_i. g_{st} is the total number of shortest paths between v_s and v_t. $(|V|-1)(|V|-2)/2$ is the total number of pairs of nodes that exist in G, excluding v_i. The *focused betweenness centrality* fBC is computed as:

$$fBC_i = \begin{cases} \frac{\sum_{v_s, v_t \in \mathbf{C}^\theta \wedge s<t} \frac{x^i_{st}}{g_{st}}}{n(n-1)/2} & v_i \notin \mathbf{C}^\theta \\ \frac{\sum_{v_s, v_t \in \mathbf{C}^\theta \wedge s<t} \frac{x^i_{st}}{g_{st}}}{(n-1)(n-2)/2} & v_i \in \mathbf{C}^\theta \end{cases} \quad .$$

Focused Random Walk Betweenness Centrality: $fRWB$. Finally, the random walk betweenness RWB in a graph $G = (V, E)$ is computed by the following steps:

1. $L = D - A$, where D is a diagonal matrix containing the degrees of the nodes and A is the adjacency matrix of G. The matrix L is called the Laplacian matrix.

2. $T_r = L_r^{-1}$, where L_r is called the reduced Laplacian. It is obtained from L by removing any single row r and the corresponding column. T_r is the reduced Laplacian's inverse.

3. The matrix T is obtained from T_r by adding a row of zeros and a column of zeros on position r.

4. RWB_i for v_i is then computed as:

$$RWB_i = \frac{\sum_{v_s, v_t \in V \wedge s<t} I_i^{(st)}}{(1/2)|V|(|V|-1)} \quad ,$$

where $I_i^{(st)}$ is the so-called intensity, from this measure's association to the current flowing through an electrical circuit [19].

$$I_i^{(st)} = 1/2 \sum_{v_j \in V} A_{ij}|T_{ij} - T_{it} - T_{js} + T_{jt}| \quad .$$

The averaging factor $(1/2)|V|(|V|-1)$ again is the number of all pairs of nodes in the graph.

For the *focused random walk betweenness* $fRWB$, we limit the computation to all paths between all pairs of seed nodes:

$$fRWB_i = \begin{cases} \frac{\sum_{v_s, v_t \in \mathbf{C}^\theta \wedge s<t} I_i^{(st)}}{(1/2)n(n-1)} & v_i \notin \mathbf{C}^\theta; \\ \frac{\sum_{v_s, v_t \in \mathbf{C}^\theta \wedge s<t} I_i^{(st)}}{(1/2)(n-1)(n-2)} & v_i \in \mathbf{C}^\theta; \end{cases} \quad .$$

The above measures fCC, fIC, fBC and $fRWB$ are the ones that we experimented with for defining the target function r, which quantifies the strength of the relation between each candidate concept and all other concepts in the topic graph $\mathbf{G}$. The graph-based labelling ranks all nodes of $\mathbf{G}$ by the chosen centrality measure and presents the top ones to the user as topic-label candidates. In the following section, we present an evaluation of the overall approach and the different centrality measures.

5. EXPERIMENTS AND EVALUATION

In this section, we describe our experiments and the results we gained on the basis of a crowd-sourcing experiment. One objective is to show the suitability of the centrality measures we propose in Section 4.1 and the differences we can observe in applying them. We discuss the chosen measures and the standard text-based method we compare to in Section 5.1. The data we used in the experiments is described in Section 5.2, including a brief analysis of the impact of removing stop URIs. Section 5.3 presents an overview of

the user study, a crucial requirement for obtaining the comparative results discussed in Section 5.4. Finally, we inspect the stability of our approach in terms of the number of seed nodes in Section 5.5.

5.1 Evaluated Methods

Pearson Correlations	fCC	fBC	fIC	$fRWB$
Degree	0.3365	0.4889	0.5072	0.6620
fCC	1	0.4432	**0.7967**	0.5118
fBC		1	0.4967	**0.8923**
fIC			1	0.6436
$fRWB$				1

Table 2: Correlation of the focused centrality measures

To keep the requirements of the user study in meaningful limits, we decided to ask the users to only evaluate fIC and $fRWB$. First, each is strongly correlated with one of the measures not evaluated and they are weakly correlated with each other. Second, by considering all paths of the topic graphs they take more information about the network topology into account than their shortest path relatives. We show the corresponding Pearson correlation coefficients in Table 2.

An important aspect is to compare our methods based on only structured data from DBpedia, to approaches that use only the documents to extract labels. We thus compare to the state-of-the-art text-based approach (TB) as described in [15]. Out of the two algorithms the authors suggest, we implemented the one for which they reported better results, the so-called "first-order relevance". The main idea is to represent candidate labels l as multinomial distribution of words $p(w|l)$. This probability represents the percentage of documents containing the word w out of the documents containing the label l. Then, a good topic label shows a distribution that is similar to the latent topic's distribution, measured using the Kullback-Leibler (KL) divergence (zero if a label perfectly matches the distribution of a topic). This value is computed as the expectation E of point-wise mutual information (PMI) between the label l and the topic words given the context D (i.e. the document corpus). The score s of a label is thus computed as:

$$s(l,\theta) = E_\theta[PMI(w,l|D)] = \sum_w (p(w|\theta)PMI(w,l|D)) \quad .$$

As in [15], for labelling we select the 1,000 most frequent noun phrases extracted from the corpus with the NLP Noun-Chunker [5] and rank them by s.

5.2 Data

For evaluating our approach and the different centrality measures, we require topics extracted and linked to DBpedia. To generate this, we ran LDA [13] on three corpora, and linked and disambiguated them using the method presented in [9]. The three corpora used are:

- The British Academic Written English Corpus (BAWE) [18] consists of 2,761 documents of proficient assessed student writing, ranging in length from about 500-5,000 words. The documents are fairly evenly distributed across four broad disciplinary areas (Arts and Humanities, Social Sciences, Life Sciences and Physical Sciences) covering 35 concrete disciplines.
- The BBC [5] corpus consists of 2,225 documents from the BBC news website corresponding to stories in five topical areas from 2004-2005: business, entertainment, politics, sport, and technology.
- The StackExchange [6] dataset consists of all discussion threads from nine forums of the StackExchange website. We chose forums that matched the general knowledge of the users participating in the user study: wordpress, webmasters, web applications, photography, gaming, game development, android, cooking and bicycles. We merged all posts of a single thread in one document and the final dataset consists of 3,709 documents, roughly 400 documents per domain on average.

We chose these three corpora because of the different text style they exhibit. We expect that the graph-based methods will be less sensitive to the text style than the text-based labelling method.

With respect to the user study, we aimed for evaluating 200 topics. Apart from the size, a user study also provides constraints by the actual user base and their background knowledge. First, topics should be understandable and coherent. To measure a topic's coherence, we used the measure published in [16] and computed as:

$$coherence(\theta; \mathbf{w}^{(\theta)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D(w_m^{(\theta)}, w_l^{(\theta)})+1}{D(w_l^{(\theta)})},$$

where $\mathbf{w}^{(\theta)} = \{w_1^{(\theta)}, .., w_M^{(\theta)}\}$ represents the set of top-M most probable words of the topic θ, $D(w)$ represents the number of documents containing the word w at least once, and $D(w, w')$ represents the number of documents containing words w and w', at least once each.

We extracted 150 topics from BAWE, 50 topics from BBC and 50 topics from StackExchange ranging from medium to high coherence. Afterwards, we manually removed 30 BAWE topics that were very specific and required domain knowledge clearly outside the expertise of our users, for example from chemistry and biology. Similar, we removed 18 BBC topics (mainly from sport and politics, which contained many names of persons that would require explicit familiarity) and 2 too technical StackExchange topics. The final 200 topics contained 120 from BAWE, 32 from BBC and 48 from StackExchange.

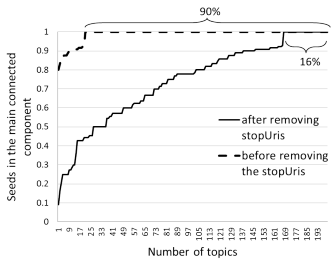


Figure 4: Seeds in the core connected component

In order to assess the impact of stop-URIs (Section 4.1), we compared topic graphs that were created before and after excluding them. We observed that when they are included in the graph, 90% of the extracted topic graphs consisted of only one connected component. This number goes down to 16% when excluding them. However, the remaining 84% of graphs contain one core component that connects on average

[5] http://opennlp.sourceforge.net/

[6] http://stackexchange.com/

69% of the seed concepts. Sense graphs that did not connect to this core remained isolated. Figure 4 shows the observed proportions.

We argue that the removal of stop-URIs results in much cleaner data. As the disambiguation algorithms have an accuracy of 70-80%[9], we have to assume 20-30% of noise among all concept seeds. When including stop-URIs, the achieved graph connectivity can be considered a 'fake' connectivity, as they bring together most of the concepts that would otherwise be isolated. For all the following experiments we therefore made sure not to extract the graph beyond the stop-URIs, and analysed only the core connected component of each topic graph.

5.3 User Study

In order to comparatively evaluate the three methods, we created a web interface to gather input from human annotators. For each randomly selected topic, annotators were given the top 5 labels produced by the three evaluated methods: TB, $fRWB$ and fIC. The labels were listed in a randomised order. The first letter of each label was capitalised so that this could not influence the users perception on the label. For each label, the annotators had to choose between: "Good Fit", "Too Broad", "Related but not a good label" and "Unrelated". There was no restriction on how many "Good Fit" labels a topic could have, so users could choose none or several. In the final data set for evaluation, each label has been annotated by exactly three different annotators. There were 54 annotators in total.

We computed the Fleiss Kappa for the inter-annotator agreement in two cases: (i) on all four classes, and (ii) on two classes obtained by collapsing "Good Fit" and "Too Broad" as well as combining "Related but not a good label" and "Unrelated". For the first case we obtained a value of 0.27, and 0.38 for the second case. These values are very much in line with the agreement obtained by [14] for the task of topic indexing. As these values correspond to the level of fair to moderate agreement, this shows that, although topic labelling is a subjective task, a certain trend in users' preferences can be observed.

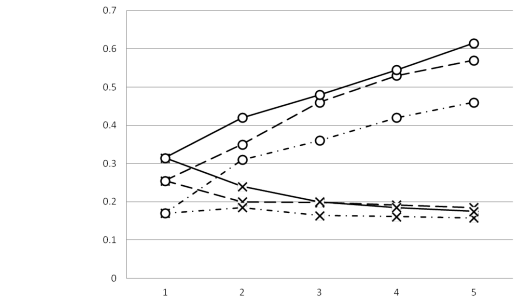
5.4 Comparative Evaluation

We evaluated two types of tests. The first one, which we call *Good Fit*, counts a Hit for a method if the recommended label was annotated as "Good Fit" by at least 2 annotators. The second type of test, called *Good-Fit-or-Broader*, counts a Hit for a method if the recommended label was annotated as "Good Fit" or as "Too Broad" by at least two annotators. This second type is aiming at a scenario of (hierarchical) classification. We expect the relation between specialised terms and general vocabulary hard to be captured using only text, but easier using structured data.

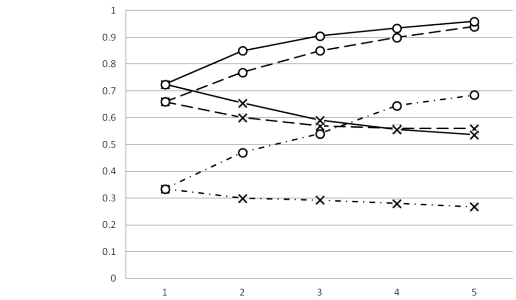
We compare the three chosen methods based on *Precision* and *Coverage*, taking the top-1 to top-5 suggested labels into account. Precision for a topic at top-k is computed as:

$$\text{Precision@}k = \frac{\#\text{Hits with rank <= k}}{k} \quad .$$

Then, we compute the average precision over all topics. As we cannot compute recall, due to the lack of ground truth, we define Coverage as the proportion of topics for which a



(a) "Good Fit"



(b) "Good Fit or Broader"

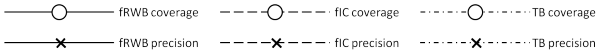


Figure 5: Precision and Coverage (y axis) @top-k (x axis) for combined corpora.

method has found at least one Hit:

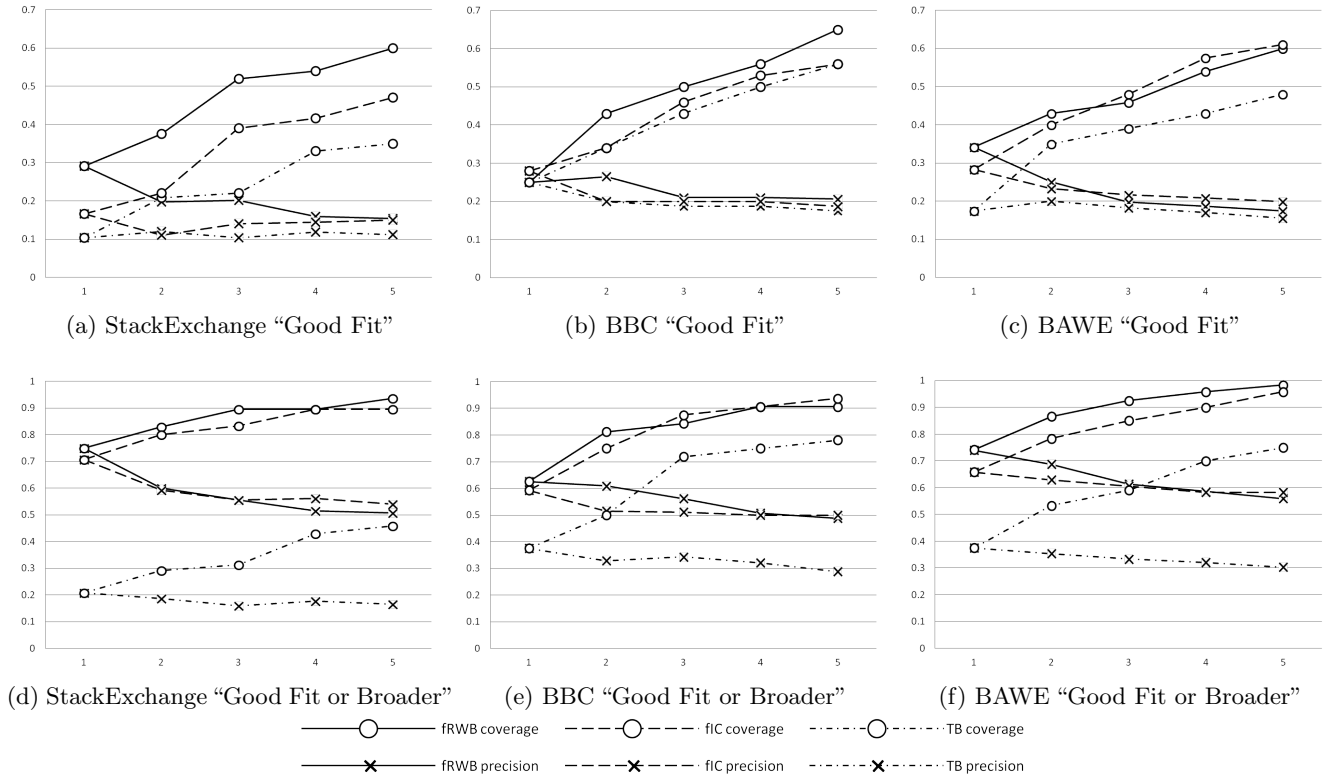
$$\text{Coverage@}k = \frac{\#\text{topics with at least one Hit at rank<=k}}{\#topics} \, .$$

Figures 5a 5b show the results for all topics combined. Figure 6 shows the results for each individual corpus.

The results indicate two advantages of our graph-based methods over the text-based one: a better coverage over all topics and a much higher ability to identify broader concepts. For the case of Good Fit, the precision values for all methods are comparable. An important difference can be seen for the precision@1 which is 31% for $fRWB$ while the text-based method achieves 17%. Regarding coverage, $fRWB$ has a Good Fit label among the top-5 in 61% of the cases, fIC in 57% and the TB in 46%.

The graph-based methods achieve significantly better results than the text-based one, in the Good-Fit-or-Broader test. In 72% of the cases the top-1 label retrieved by $fRWB$ was either a Good Fit or a Too Broad label. fIC scores 0.66 and TB 0.335. This shows that our approach is better suited for a classification scenario. This also confirms the intuition that the text-based labelling methods encounter problems identifying broader terms. As for coverage on all corpora, $fRWB$ achieves 96% in top-5, while fIC covers 94% and TB 68%.

The analysis of the different corpora provides interesting insights also. Particularly the StackExchange fora corpus highlights differences. All three methods have their worst precision on the Good Fit test on this corpus, being almost constantly under 20%. As expected, this corpus poses problems especially for the text-based method, whose coverage@5 in the Good Fit test is 0.35, with $fRWB$ scoring 0.6. On the same corpus, in the Good-Fit-or-Broader test, TB



(a) StackExchange "Good Fit" (b) BBC "Good Fit" (c) BAWE "Good Fit"

(d) StackExchange "Good Fit or Broader" (e) BBC "Good Fit or Broader" (f) BAWE "Good Fit or Broader"

Figure 6: Precision and Coverage (y axis) @top-k (x axis) for the three corpora.

has a coverage@5 of 45% whereas the $fRWB$ scores 93% and fIC 90%. Regarding the Good-Fit-or-Broader test on each corpus, the coverage@5 of $fRWB$ and fIC reaches more than 90%. More variation is seen in the converage@5 of the TB method, which is 78% on the BBC corpus, slightly lower on the BAWE corpus, while on StackExchange it results in its worst coverage@5 of less than 50%.

These results show that the graph-based methods on DBpedia can achieve better results than the standard text-based methods. The text-based method is also more sensitive to the type of text. The graph-based methods are able to retrieve better labels without a high drop in quality for forum text. The biggest difference is observed in their bias towards broader labels as compared to the text-based method. More experiments are needed with other knowledge bases than only DBpedia in order to conclude if the bias towards broader labels is due to the nature of graph-based measures or due to the nature of concepts in DBpedia. However, the results indicate that the graph-based labelling is more suited for recommendation scenarios where a good coverage is more important than a good precision.

5.5 Stability of Graph Measures

Topic labelling using external knowledge strongly depends on the quality of the linking of topic words. In our experiments, the disambiguation algorithm received the top 15 words of each topic. Usually, there are topic terms that cannot be linked, because they do not have a corresponding DBpedia concept. Moreover, we also want to support cases when the input topics are not necessarily probabilistic latent topics, for instance if they are extracted from a phrase, and contain very few words. Therefore, we analyse the impact of the number of disambiguated concepts. We achieve this by inspecting the number of concepts in core connected component of the topic graph.

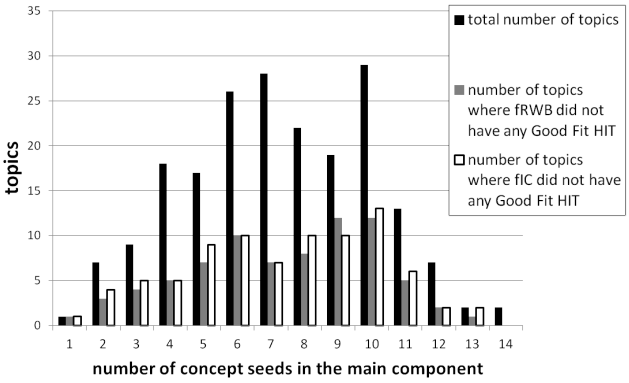


Figure 7: Influence of number of seed nodes

We selected the topics for which the graph-based methods did not find any label annotated with Good Fit by at least two annotators. Then, we statistically determined if the number of core concepts in these cases is biased in comparison to all topics. The distributions are shown in Figure 7. For each method, we computed the Chi Square goodness of fit statistic with respect to the distribution of all topics. In both cases, there was no significant difference between the mistaken topics distribution and the distribution of all topics. For $fRWB$ we obtained $\chi^2(13, n = 77) = 7.10, p > 0.10$, and for fIC we obtained $\chi^2(13, n = 85) = 7.44, p > 0.10$.

This result has an important practical significance, as it shows that even with less than 5 core concepts the labelling can be as successful as with more than 5 or even more than 10 core concepts. We also analysed with how many seed concepts the different centrality measures converge to the final set of top-5 labels. We noticed that for all measures, the first 5 concept seeds already established at least 2 labels of the final top-5 set. We also observed that fCC is not very sensitive to new concepts once it identified concepts very close to its seed concepts, while fBC and $fRWB$ are most sensitive to each individual seed concept.

6. CONCLUSION AND FUTURE WORK

In this work, we investigated approaches for graph-based topic labelling using DBpedia. We extract the DBpedia subgraph of topic concepts and adapt network centrality measures to identify concepts that promise to be good labels for a topic. On the basis of a crowd-sourcing experiment, we showed that the graph-based approaches perform constantly better than a state-of-the-art text-based method. The most important improvements are (i) better corpus coverage, and (ii) much higher ability to identify broader labels. We envisage applications that support users in the tasks of topic-labelling and navigation – either by recommending a set of top labels or by recommending exploration directions.

However, none of these approaches is yet ready for fully automated labelling. In this perspective, we continue our research by investigating graph patterns (e.g., density of the topic graph) that could identify particular centrality measures suited in particular situations.

Linking topics from a corpus to external knowledge bases like DBpedia has more benefits than just topic labelling. For example, relations and similarities between different topics can be identified based on the graph overlap between topics. The extent to which topics overlap or are similar to one another can help the user assess the suitability of the chosen number of topics. Finally, a network of topics is obtained for a corpus that can serve as basis for corpus navigation. There are many interesting research directions in the area of graph mining to topic / document analysis, and the work presented here is barely scratching the surface.

7. ACKNOWLEDGEMENTS

This work was jointly supported by Science Foundation Ireland (SFI) partly under Grant No. 08/CE/I1380 (Lion-2) and partly under Grant No. 08/SRC/I1407 (Clique: Graph and Network Analysis Cluster), and by the European Union (EU) under grant no. 257859 (ROBUST integrating project).

8. REFERENCES

[1] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. Dbpedia-a crystallization point for the web of data. *JWS*, 7(3):154–165, 2009.

[2] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[3] D. Carmel, H. Roitman, and N. Zwerdling. Enhancing cluster labeling using wikipedia. In *SIGIR '09*, pages 139–146, 2009.

[4] K. Coursey, R. Mihalcea, and W. Moen. Using encyclopedic knowledge for automatic topic identification. In *CoNLL '09*, pages 210–218, 2009.

[5] D. Greene and P. Cunningham. Practical solutions to the problem of diagonal dominance in kernel document clustering. In *ICML'06*, pages 377–384, 2006.

[6] J. Hoffart, F. M, S. K. Berberich, G. Weikum, and I. Saclay. Yago2: A spatially and temporally enhanced knowledge base from wikipedia. *Commun. ACM*, 2009.

[7] T. Hofmann. Probabilistic latent semantic analysis. In *Uncertainty in Artificial Intelligence (UAI '99)*, pages 289–296, 1999.

[8] http://dublincore.org/. DCMI: The Dublin Core Metadata Initiative, 2012. [accessed 07-August-2012].

[9] I. Hulpuş, C. Hayes, M. Karnstedt, and D. Greene. An eigenvalue based measure for word-sense disambiguation. In *FLAIRS 2012*, 2012.

[10] J. H. Lau, K. Grieser, D. Newman, and T. Baldwin. Automatic labelling of topic models. In *ACL: Human Language Technologies*, pages 1536–1545, 2011.

[11] W. Li and A. McCallum. Pachinko allocation: Dag-structured mixture models of topic correlations. In *ICML '06*, pages 577–584, 2006.

[12] D. Magatti, S. Calegari, D. Ciucci, and F. Stella. Automatic labeling of topics. In *Intelligent Systems Design and Applications*, pages 1227–1232, 2009.

[13] A. K. McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

[14] O. Medelyan, I. H. Witten, and D. Milne. Topic indexing with wikipedia. In *Workshop on Wikipedia and Artificial Intelligence (WIKIAI '08)*, 2010.

[15] Q. Mei, X. Shen, and C. Zhai. Automatic labeling of multinomial topic models. In *SIGKDD '07*, pages 490–499, 2007.

[16] D. M. Mimno, H. M. Wallach, E. M. Talley, M. Leenders, and A. McCallum. Optimizing semantic coherence in topic models. In *EMNLP '11*, pages 262–272, 2011.

[17] M. Muhr, R. Kern, and M. Granitzer. Analysis of structural relationships for hierarchical cluster labeling. In *SIGIR '10*, pages 178–185, 2010.

[18] H. Nesi, S. Gardner, P. Thompson, and P. Wickens. British academic written english (bawe) corpus. *Universities of Warwick, Reading and Oxford Brookes, under funding from ESRC (RES-000-23-0800)*, 2007.

[19] M. Newman. A measure of betweenness centrality based on random walks. *Social networks*, 27:39–54, 2005.

[20] M. Newman. *Networks. An Introduction*. Oxford University Press, 2010.

[21] T. Nomoto. Wikilabel: an encyclopedic approach to labeling documents en masse. In *CIKM '11*, pages 2341–2344, 2011.

[22] K. Stephenson and M. Zelen. Rethinking centrality: Methods and examples. *Social Networks*, 11:1 – 37, 1989.

[23] Z. Syed, T. Finin, and A. Joshi. Wikipedia as an ontology for describing documents. In *ICWSM '08*, 2008.

[24] P. Treeratpituk and J. Callan. Automatically labeling hierarchical clusters. In *Conference on Digital Government Research*, pages 167–176, 2006.

[25] W3C. SKOS: Simple Knowledge Organization System. `http://www.w3.org/2004/02/skos/`, 2009.