Extended Document Representation for Search Result Clustering*

S. Hoa Nguyen, Wojciech Świeboda, and Grzegorz Jaśkiewicz

Faculty of Mathematics, Informatics, and Mechanics, The University of Warsaw
Banacha 2, 02-097 Warsaw, Poland
hoa@mimuw.edu.pl, wswieb@mimuw.edu.pl, jaskiewicz@mimuw.edu.pl

Abstract. Organizing query results into clusters facilitates quick navigation through search results and helps users to specify their search intentions. Most meta-search engines group documents based on short fragments of source text called *snippets*. Such a model of data representation in many cases shows to be insufficient to reflect semantic correlation between documents. In this paper, we discuss a framework of document description extension which utilizes domain knowledge and semantic similarity. Our idea is based on application of Tolerance Rough Set Model, semantic information extracted from source text and domain ontology to approximate concepts associated with documents and to enrich the vector representation.

Keywords: Text mining, semantic clustering, DBpedia, document grouping, PubMed, bibliometric measure.

1 Introduction

Although the performance of search engines is improving every day, searching the Web can be a tedious and time-consuming task because (1) search engines can index only a part of the "indexable Web" due to its huge size and a highly dynamic nature, and (2) the user's intention is not clearly expressed in general, short queries. In effect, a search engine may return as much as hundreds of thousands of relevant documents. One approach to manage the large number of results is clustering. The concept arises from document clustering in information retrieval domain: find a grouping for a set of documents so that documents belonging to the same cluster are similar and documents belonging to different clusters are dissimilar. Search results clustering can thus be defined as a process of automatic grouping of search results into thematic groups and discovering concise descriptions of these groups. Clustering of search results can help the

* The authors are supported by the grant N N516 077837 from the Ministry of Science and Higher Education of the Republic of Poland and by the National Centre for Research and Development (NCBiR) under Grant No. SP/I/1/77065/10 by the strategic scientific research and experimental development program: "Interdisciplinary System for Interactive Scientific and Scientific-Technical Information".

R. Bembenik et al. (Eds.): Intelligent Tools for Building a Scient. Info. Plat., SCI 390, pp. 77–95.
springerlink.com

user navigate through a large set of documents more efficiently and help users specify their intentions. By providing concise, accurate descriptions of clusters, it lets users localize interesting documents faster. Most notably, document clustering algorithms are designed to work on relatively large collections of full-text documents, as opposed to search results clustering algorithms, which are supposed to work on a set of short text descriptions (usually between 10 and 20 words).

The earliest works on clustering results were done by Hearst and Pedersen on scather/gather system [3], followed by an application to Web documents and search results by Zamir and Etzioni [16] to create groups based on novel algorithm suffix tree clustering. Inspired by their work, a Carrot framework was created by Weiss [2,14] to facilitate research on clustering search results. This has encouraged others to contribute new clustering algorithms under the Carrot framework like LINGO [6] and AHC [15]. The main problem occurred in all mentioned works is the fact that many snippets remain unrelated with a genuine content of documents because of their short representation.

Several works in the past have been devoted to the problem of document description enrichment. In [5] a method of snippet extension was investigated. The main idea was based on application of collocation similarity measure to enrich the vector representation. In [12], the author presented a method of associating terms in document representation with similar concepts drawn from domain ontology.

In this paper, we present a generalized scheme for the problem mentioned above. We investigate two levels of extensions. The first one is related to extending a concept space for data representation and the second one is related to associating terms in the concept space with semantically related concepts. The first extension level is performed by incorporating semantic information extracted from a document content (such as citations) or from document meta-data (like authors, conferences). The second one is achieved by application of Tolerance Rough Set Model [11] and domain ontology, in order to approximate concepts existing in document description and to enrich the vector representation of the document.

The paper is organized as follows. In the second section we present a framework of **T**olerance **R**ough **S**et **M**odel (TRSM). In Section 3 we discuss the application of *Generalized* TRSM to enriching document descriptions, whereas section 4 is devoted to experiments, followed by conclusions in section 5.

2 Generalized Approximation Space and TRSM

Rough set theory was originally developed [7] as a tool for data analysis and classification. It has been successfully applied in various tasks, such as feature selection/extraction, rule synthesis and classification [4]. In this chapter we will present fundamental concepts of rough sets with illustrative examples. Some extensions of the Rough Set model are described, concentrating on the use of rough sets to synthesize approximations of concepts from data.

Consider a non-empty set of object U called the universe. Suppose we want to define a concept over the universe of objects U. Let us assume that the concept can be represented as a subset X of U. The central point of Rough Set theory is the notion of set approximation: any set in U can be approximated by its *lower* and *upper approximation*.

2.1 Generalized Approximation Spaces

The classical Rough Set theory is based on equivalence relation that divides the universe of objects into disjoint classes. By definition, an equivalence relation $R \subseteq U \times U$ is required to be reflexive, symmetric, and transitive. Practically, in some applications, the requirement for equivalent relation has shown to be too strict. Concepts arising in many domains are by their nature imprecise and overlapping eachother. For example, let us consider a collection of scientific documents and keywords describing those documents. It is clear that each document can be assigned several keywords and a keyword can be associated with many documents. Thus, in the universe of documents, keywords can form overlapping classes.

Skowron [11] has introduced a generalized tolerance space by relaxing the relation R to a tolerance relation, where transitivity property is not required. Formally, the generalized approximation space is defined as a quadruple $\mathcal{A} = (U, I, \nu, P)$, where

1. U is a non-empty set of objects (*a universe*).
2. $I : U \to \mathcal{P}(U)$ is an *uncertainty function* ($\mathcal{P}(U)$ is a set of all subsets of U) satisfying conditions:
   - $x \in I(x)$ for $x \in U$
   - $y \in I(x) \iff x \in I(y)$, for any $x, y \in U$.

   Thus, the relation $xRy \iff y \in I(x)$ is a *tolerance relation* (i.e. reflexive, symmetric) and $I(x)$ is a *tolerance class* of x.
3. $\nu : \mathcal{P}(U) \times \mathcal{P}(U) \to [0, 1]$ is a *vague inclusion function*.
   Vague inclusion ν is a kind of membership function but extended to functions over $\mathcal{P}(U) \times \mathcal{P}(U)$ to measure the degree of inclusion between two sets. Vague inclusion must be *monotonic* with respect to the second argument, i.e., if $Y \subseteq Z$ then $\nu(X, Y) \leq \nu(X, Z)$ for $X, Y, Z \subseteq U$.
4. $P : I(U) \to \{0, 1\}$ is a *structurality function*.
   The introduction of structurality function $P : I(U) \to \{0, 1\}$ ($I(U) = \{I(x) : x \in U\}$) allows us to enforce additional global conditions on sets $I(x)$ considered to be approximated. In generation of approximations, only sets $X \in I(U)$ for which $P(X) = 1$ (referred to as *P-structural element* in U) are considered. For example, a function $P_\alpha(X) = 1 \iff |X|/|U| > \alpha$ will discard all subsets that are relatively smaller than certain percentage (given by α) of U.

Together with uncertainty function I, vague inclusion function ν defines the *rough membership function* for $x \in U, X \subseteq U$:

$$\mu_{I,\nu}(x, X) = \nu(I(x), X)$$

Lower and upper approximations in $\mathcal{A}$ of any $X \subseteq U$ are then defined as

$$\mathbf{L}_{\mathcal{A}}(X) = \{x \in U : P(I(x)) = 1 \ \wedge \nu(I(x), X) = 1\} \quad (1)$$

$$\mathbf{U}_{\mathcal{A}}(X) = \{x \in U : P(I(x)) = 1 \ \wedge \nu(I(x), X) > 0\} \quad (2)$$

With the definition given above, generalized approximation spaces can be used in any application where I, ν and P are appropriately determined.

2.2 Tolerance Rough Set Model

Tolerance Rough Set Model (TRSM) was developed [10,13] as a basis to model documents and terms in Information Retrieval, Text Mining, etc. With its ability to deal with vagueness and fuzziness, Tolerance Rough Set Model seems to be a promising tool to model relations between terms and documents. In many Information Retrieval problems, especially in document clustering, defining the relation (i.e. similarity or distance) between document-document, term-term or term-document is essential. In Vector Space Model, is has been noticed [13] that a single document is usually represented by relatively few terms[1]. This results in zero-valued similarities which decreases quality of clustering. The application of TRSM in document clustering was proposed as a way to enrich document and cluster representation with the hope of increasing clustering performance.

The idea is to capture conceptually related index terms into classes. For this purpose, the tolerance relation R is determined as the co-occurrence of index terms in all documents from D. The choice of co-occurrence of index terms to define tolerance relation is motivated by its meaningful interpretation of the semantic relation in context of IR and its relatively simple and efficient computation.

Let $D = \{d_1, \ldots, d_N\}$ be a set of documents and $T = \{t_1, \ldots, t_M\}$ set of *index terms* for D. With the adoption of Vector Space Model [1], each document d_i is represented by a weight vector $[w_{i1}, \ldots, w_{iM}]$ where w_{ij} denotes the weight of term t_j in document d_i. TRSM is an approximation space $\mathcal{R} = (T, I_\theta, \nu, P)$ determined over the set of terms T as follows:

- **Uncertainty Function:** The parameterized uncertainty function I_θ is defined as
$$I_\theta(t_i) = \{t_j \mid f_D(t_i, t_j) \geq \theta\} \cup \{t_i\}$$
where $f_D(t_i, t_j)$ denotes the number of documents in D that contain both terms t_i and t_j and θ is a parameter set by an expert.
The set $I_\theta(t_i)$ is called the *tolerance class* of index term t_i.
- **Vague Inclusion Function:** To measure degree of inclusion of one set in another, the vague inclusion function is defined as is defined as
$$\nu(X, Y) = \frac{|X \cap Y|}{|X|}$$
It is clear that this function is monotone with respect to the second argument.

[1] In other words, the number of non-zero values in document's vector is much smaller than vector's dimension – the number of all index terms.

Table 1. Tolerance classes of terms generated from 200 snippets return by Google search engine for a query "jaguar" with $\theta = 9$;

Term	Tolerance class	Document frequency
Atari	**Atari, Jaguar**	**10**
Mac	**Mac, Jaguar, OS, X**	**12**
onca	onca, Jaguar, Panthera	9
Jaguar	Atari, Mac, onca, Jaguar, club, Panthera, new, information, OS, site, Welcome, X, Cars	185
club	Jaguar, club	27
Panthera	**onca, Jaguar, Panthera**	**9**
new	Jaguar, new	29
information	Jaguar, information	9
OS	Mac,Jaguar, OS, X	15
site	Jaguar, site	19
Welcome	Jaguar, Welcome	21
X	Mac, Jaguar, OS, X	14
Cars	**Jaguar, Cars**	**24**

– **Structural Function:** All tolerance classes of terms are considered as structural subsets: $P(I_\theta(t_i)) = 1$ for all $t_i \in T$.

The membership function μ for $t_i \in T$, $X \subseteq T$ is then defined as $\mu(t_i, X) = \nu(I_\theta(t_i), X) = \frac{|I_\theta(t_i) \cap X|}{|I_\theta(t_i)|}$ and the lower and upper approximations of any subset $X \subseteq T$ can be determined – with the obtained tolerance $\mathcal{R} = (T, I, \nu, P)$ – in the standard way

$$\mathbf{L}_\mathcal{R}(X) = \{t_i \in T \mid \nu(I_\theta(t_i), X) = 1\}$$
$$\mathbf{U}_\mathcal{R}(X) = \{t_i \in T \mid \nu(I_\theta(t_i), X) > 0\}$$

2.3 Example

Consider a universe of unique terms extracted from a set of search result snippets returned from Google search engine for a "famous" query: **jaguar**, which is frequently used as a test query in information retrieval because it is a polysemy, i.e., a word that has several meanings, especially in the Web. The word jaguar can have the following meanings:

– jaguar as a cat (*panthera onca* - http://dspace.dial.pipex.com/agarman/jaguar.htm);
– jaguar as a car;
– jaguar was a name for a game console made by Atari - http://www.atari-jaguar64.de;
– it is also a codename for Apple's newest operating system MacOS X - http://www.apple.com/macosx.

Tolerance classes are generated for threshold $\theta = 9$. It is interesting to observe (Table 1) that generated classes reveal different meanings of the word "jaguar": a cat, a car, a game console, an operating system and some more.

In the context of Information Retrieval, a tolerance class represents a concept that is characterized by terms it contains. By varying the threshold θ , one can control the degree of relatedness of words in tolerance classes (or the preciseness of the concept represented by a tolerance class).

One interpretation of given approximations is as follows: if we treat X as a concept described vaguely by index terms it contains, then $\mathbf{U}_{\mathcal{R}}(X)$ is the set of concepts that share some semantic meanings with X, while $\mathbf{L}_{\mathcal{R}}(X)$ is a "core" concept of X.

3 Applications of TRSM in Semantic Search

Tolerance Rough Set Model was applied to many text mining problems, including document clustering, search result clustering [5,10], and automatic syllabus generation. One can list the three most important applications of TRSM in text mining:

- Enriching document representation;
- Extended weighting scheme;
- TRSM based clustering algorithms.

In this paper we propose a new application of TRSM in Semantic Search, one actually developed within SYNAT project. The idea is to extend the representation of documents (or snippets) by additional semantic information extracted from text sources and/or from meta-data like citations, authors, publishers, publication years and semantic concepts related to the document. Three Tolerance Rough Set models will be discussed in this section: standard TRSM, extended TRSM by citations and extended TRSM by sematic concepts. We also present, how to apply these models to enrich documents' representation.

3.1 Standard TRSM

Let $D = \{d_1, \ldots, d_N\}$ be a set of documents and $T = \{t_1, \ldots, t_M\}$ the set of *index terms* for D. The tolerance rough set model for term space was described in previous Section.

$$\mathcal{R}_0 = (T, I_\theta, \nu, P)$$

In this model a document, which is associated with a *bag of words/terms*, is represented by its upper approximation, i.e. the document $d_i \in D$ is represented by

$$\mathbf{U}_{\mathcal{R}}(d_i) = \{t_i \in T \mid \nu(I_\theta(t_i), d_i) > 0\}$$

The extended weighting scheme is inherited from the standard TF-IDF by:

$$w^*_{ij} = \begin{cases} (1 + \log f_{d_i}(t_j)) \log \frac{N}{f_D(t_j)} & \text{if } t_j \in d_i \\ 0 & \text{if } t_j \notin \mathbf{U}_{\mathcal{R}}(d_i) \\ \min_{t_k \in d_i} w_{ik} \frac{\log \frac{N}{f_D(t_j)}}{1 + \log \frac{N}{f_D(t_j)}} & \text{otherwise} \end{cases}$$

Table 2. Example snippet and its two vector representations in standard TRSM

Title: EconPapers: Rough sets bankruptcy prediction models versus auditor
Description: Rough sets bankruptcy prediction models versus auditor signalling rates. Journal of Forecasting, 2003, vol. 22, issue 8, pages 569-586. Thomas E. McKee. ...

Original vector		**Enriched vector**	
Term	Weight	Term	Weight
auditor	0.567	auditor	0.564
bankruptcy	0.4218	bankruptcy	0.4196
signalling	0.2835	signalling	0.282
EconPapers	0.2835	EconPapers	0.282
rates	0.2835	rates	0.282
versus	0.223	versus	0.2218
issue	0.223	issue	0.2218
Journal	0.223	Journal	0.2218
MODEL	0.223	MODEL	0.2218
prediction	0.1772	prediction	0.1762
Vol	0.1709	Vol	0.1699
		applications	0.0809
		Computing	0.0643

The extension ensures that each term occurring in the upper approximation of d_i but not in d_i itself has a weight smaller than the weight of any terms in d_i. Normalization by vector's length is then applied to all document vectors:

$$w_{ij}^{new} = \frac{w_{ij}^*}{\sqrt{\sum_{t_k \in d_i} (w_{ij}^*)^2}}$$

The example of standard TRSM is presented in Table 2.

3.2 Extended TRSM by Citation

Let $D = \{d_1, \ldots, d_N\}$ be a set of documents and $T = \{t_1, \ldots, t_M\}$ the set of *index terms* for D. Let $B = \{b_1, \ldots, b_K\}$ be the set of bibliography items that are cited by documents from D.

The extended tolerance rough set model for terms and citations is a pair:

$$\mathcal{R}_1 = (\mathcal{R}_T, \mathcal{R}_B)$$

where

$\mathcal{R}_T = (T, I_{\theta_T}, \nu, P)$ and $\mathcal{R}_B = (B, I_{\theta_B}, \nu, P)$ are TRSM defined for term space T and bibliography item space B, respectively.

In the extended model, each document $d_i \in D$ is associated with a pair (T_i, B_i), where T_i is the set of terms that occur in d_i and B_i is the set of

bibliography items cited by d_i. Each document can be represented by a pair of upper approximations, i.e.,

$$d_i \dashrightarrow (\mathbf{U}_{\mathcal{R}_T}(d_i), \mathbf{U}_{\mathcal{R}_B}(d_i))\}$$

where

$$\mathbf{U}_{\mathcal{R}_T}(d_i) = \{t_i \in T \mid \nu(I_{\theta_T}(t_i), d_i) > 0\}$$
$$\mathbf{U}_{\mathcal{R}_B}(d_i) = \{b_i \in B \mid \nu(I_{\theta_B}(b_i), d_i) > 0\}$$

Once can observe some properties of the extended model.

- if $B = \emptyset$, the extended model $\mathcal{R}_1$ is the standard TRSM.
- if $B = \emptyset$ and $\theta_T = card(D) + 1$, documents in D are represented by their original text without information about citations.
- if $T = \emptyset$ and $\theta_B = card(D) + 1$, documents in D are represented by bibliography items occurred in the documents without information about an original text.
- if $\theta_T = \theta_B = card(D) + 1$, each document $d_i \in D$ is represented by an original text and bibliography items cited by d_i.

In this paper we are investigating the last three cases and their influence to the search result clustering problem.

3.3 Extended TRSM by Semantic Concepts

Let $D = \{d_1, \ldots, d_N\}$ be a set of documents and $T = \{t_1, \ldots, t_M\}$ the set of *index terms* for D. Let $B = \{b_1, \ldots, b_K\}$ be the set of bibliography items that are cited by documents from D and let C be the set of concepts from a given domain knowledge (e.g. the concepts from DBpedia).

The extended tolerance rough set model based on both citations and semantic concepts is a tuple:

$$\mathcal{R}_C = (\mathcal{R}_T, \mathcal{R}_B, \mathcal{R}_C, \alpha_n)$$

where

$\mathcal{R}_T$, $\mathcal{R}_B$ and $\mathcal{R}_C$ are tolerance spaces determined over the set of terms T, the set of bibliography items B and the set of concepts in the knowledge domain C, respectively.

$\alpha_n : \mathcal{P}(T) \longrightarrow \mathcal{P}(C)$ is called the *semantic association* for terms. For any $T_i \subset T$, $\alpha_n(T_i)$ is the set of top n most associated concepts for T_i, see [12].

In this model, each document $d_i \in D$ associated with a pair (T_i, B_i) is represented by a triple:

$$d_i \dashrightarrow (\mathbf{U}_{\mathcal{R}_T}(d_i), \mathbf{U}_{\mathcal{R}_B}(d_i), \alpha_n(T_i))\}$$

In Figure 1 we can see an example of a biomedical article and the list of concepts associated with the article.

BMC Psychiatry

BioMed Central

Research article

Open Access

Cognitive function during early abstinence from opioid dependence: a comparison to age, gender, and verbal intelligence matched controls

Pekka Rapeli*[1,2,7], Reetta Kivisaari[3], Taina Autti[3], Seppo Kähkönen[4,5], Varpu Puuskari[6], Olga Jokela[1] and Hely Kalska[7]

Abstract

Background: Individuals with opioid dependence have cognitive deficits during abuse period in attention, working memory, episodic memory, and executive function. After protracted abstinence consistent cognitive deficit has been found only in executive function. However, few studies have explored cognitive function during first weeks of abstinence. The purpose of this study was to study cognitive function of individuals with opioid dependence during early abstinence. It was hypothesized that cognitive deficits are pronounced immediately after peak withdrawal symptoms have passed and then partially recover.

personality disorder was more common than known frequencies of these disorders among substance abusers. This also should be taken into account when comparing our results with other studies. The group matching procedure of our study was based on VIQ whereas in most other studies the matching is based on education. Though matching for VIQ or premorbid IQ is not totally exceptional in opioid abuse studies [6,11], this should be taken into account when comparing our results with other studies. Finally, the suggestion for transient cognitive deficit due to high stress system activation is very preliminary. Longitudinal studies with relevant cognitive and neuroendocrine variables may determine if suggested associations between these variables exist.

Conclusion

Fluid intelligence, working memory, and executive function deficit observed in opioid withdrawal patients during

References

1. Kamboj SK, Tookman A, Jones L, Curran HV: **The effects of immediate-release morphine on cognitive functioning in patients receiving chronic opioid therapy in palliative care.** *Pain* 2005, 117:388-395.
2. Sjogren P, Christrup LL, Petersen MA, Hojsted J: **Neuropsychological assessment of chronic non-malignant pain patients treated in a multidisciplinary pain centre.** *European J Pain* 2005, 9:453-462.
3. Chapman SL, Byas-Smith MG, Reed BA: **Effects of intermediate- and long-term use of opioids on cognition in patients with chronic pain.** *Clin J Pain* 2002, 18:383-390.
4. Grant I, Adams KM, Carlin AS, Rennick PM, Judd LL, Schoof K: **The collaborative neuropsychological study of polydrug users.** *Arch Gen Psychiatry* 1978, 35:1063-1074.
5. Roumeville BJ, Jones C, Novelly RA, Kleber H: **Neuropsychological functioning in opiate addicts.** *J Nerv Ment Dis* 1982, 170:209-216.
6. Ornstein TJ, Iddon JL, Baldacchino AM, Sahakian BJ, London M, Everitt BJ, Robbins TW: **Profiles of cognitive dysfunction in chronic amphetamine and heroin abusers.** *Neuropsychopharmacology* 2000, 23:113-126.
7. Rogers RD, Everitt BJ, Baldacchino A, Blackshaw AJ, Swainson R, Wynne K, Baker NB, Hunter J, Carthy T, Booker E, London M, Deakin JFW, Sahakian BJ, Robbins TW: **Dissociable deficits in the**

Top 20 concepts:
"Opioid antagonist" "Neurocognitive" "Opiate replacement therapy" "Paced Auditory Serial Addition Test" "Withdrawal" "Opioid-induced hyperalgesia" "Methadone" "Benzodiazepine withdrawal syndrome" "Opioid" "7-Hydroxymitragynine" "Nalmefene" "DAMGO" "Cyprenorphine" "RB-101" "Meptazinol" "Post acute withdrawal syndrome" "IC-26" "Euphoriant" "Cognitive deficit" "Cognitive neuropsychology"

Fig. 1. An example of an article and the list of top 20 concepts that are related to the article

4 Case Study: Search Result Clustering of Biomedical Articles

The extended TRSM models described in the previous section are applied to enrich descriptions of biomedical articles from the PubMed Central database [9]. The purpose of this section is to evaluate document representation models. In experiments we investigate the following document representations:

- Abstract based representation.
- Citation based representation.
- Semantic concept based representation.
- Abstract enriched by citation.

These models are applied to a *search result clustering problem*. By analyzing cluster quality one can evaluate document representation models.

4.1 Data Sources

PubMed Central[9] (PMC) is a free online archive of journal articles in biomedicine and life sciences. A subset of this database, PMC Open Access subset, consists of articles available under Creative Commons license or similar, thus may be downloaded in bulk from PMC. This subset contains 200000 articles (roughly 10% of the PMC database) and provides a base text corpus for our experiments, further restricted by specific search queries. All articles are provided

along with rich metadata from MEDLINE database, and most articles in MEDLINE/PubMed database are indexed with MeSH (Medical Subject Headings), a controlled vocabulary. MeSH terms are assigned to texts by subject experts.

The second data source that we utilize in our experiments, one which provides an alternative document representation is DBpedia. The interested reader will find further details in [12], but for the purpose of this article it suffices to think of an abstract module which takes as the input an article and provides a list of DBpedia entries associated with this article, along with degrees of association.

4.2 Experiment Set-Up

The aim of our experiments is to explore clusterings induced by different document representations (lexical, semantic and structural). To cluster documents we adopt an algorithm LINGO in a Carrot clustering library [6]. The main idea of the algorithm is to apply a Singular Value Decomposition (SVD) method to document indexing. One of the advantages of LINGO is the ability to assign labels to clusters.

The diagram of our experiments is shown in Figure 2. An experiment path (from querying to search result clustering) consists of three stages:

- Search and filter documents matching to a query. Documents in a search result list are represented by *snippets* and/or *titles*.
- Extend representations of the documents by *citations* and/or *semantically similar concepts* from DBpedia.
- Cluster the document search results.

At this point, we limit our exposure in experimental section to a query: "*hippocampus cortisol*".

This query returned 989 papers from PubMed Central (PMC) search engine. 196 of these documents are available in PubMed Central Open Archive subset (as of this writing). We have further restricted this set to 134 journal articles which were suitable for our investigations (i.e. full text was available in Archiving and Interchange Tag Set format).

We will treat LINGO as a clustering engine. Conceptually, the algorithm takes as an input information about documents in terms of pairs (*title*, *snippet*).

While Carrot^2 workbench application is integrated with PubMed search engine, we use a Web based search application which directly queries PMC (rather than PubMed) database. For PubMed queries in Carrot, "snippet" field used by the system contains *document abstract* (along with basic document meta-data). Hence, a document representation which induces a grouping that serves as a natural benchmark is one which places document abstracts in "snippet" field. Please keep in mind that the set of PubMed documents clustered directly by Carrot^2 workbench and those clustered during our investigations differ, as we are limited to PMC Open Archive subset.

We have investigated various document representations: based on abstracts, based on citations (other articles referenced in bibliography as well as articles

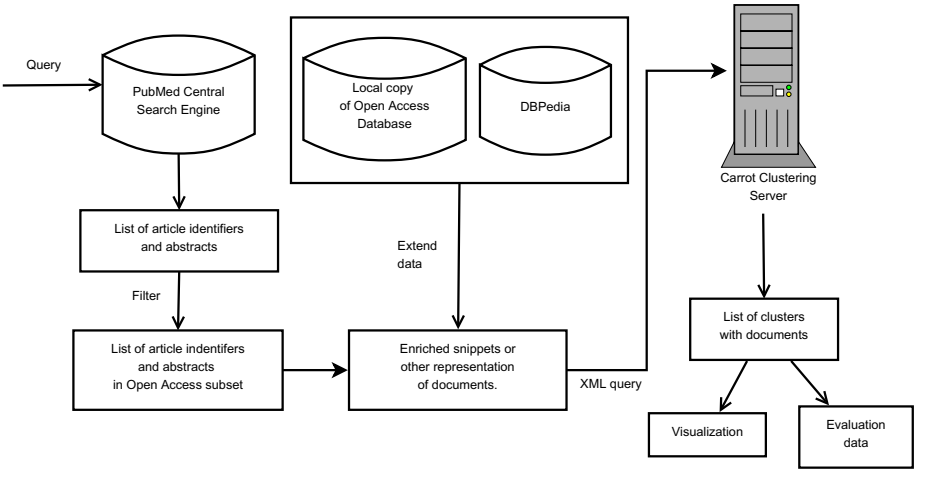

Fig. 2. Experiment diagram

that reference a given paper) and based on DBpedia entries semantically similar to a given document [12]. We have also considered mixed representations (e.g. abstracts augmented with citations) and experimented with varying number of DBpedia entries used in document representation.

Despite a slightly broader coverage of our work, on most graphs that follow, for the purpose of clarity we limit our exposure to selected representative document representations (and hence clusterings).

4.3 Results of Experiments

The subsequent subsections will focus on the following analyses:

- First, we will take a look at an example cluster that was only discovered when document representation was enriched with information about citations.
- Secondly, we will briefly look at stability of resulting clusters with respect to document representation. How does a clustering change if a baseline representation (based on abstracts) is enriched by including citations? How does it change when the representation is replaced with one based on DBpedia concepts?
- Next, we will see whether there is any common information in disjoint document representations.
- We will perform validation of clustering results using MeSH terms associated with articles.
- Finally, we will take a look at certain structural properties of resulting clusterings.

4.4 Example Cluster

Table 3 shows an example cluster (labeled “Body Weight’) discovered after extending baseline document representation with citation information. In

Table 3. A cluster labeled "Body Weight", discovered after baseline document representation was extended with citation information. Column "Grouping (abstract)" shows original (baseline) groups assigned to each document (two of them were previously unassigned to any group), whereas the third column lists MeSH terms associated with each document (these terms were unavailable for the fourth document). We have emphasised concepts that seem (subjectively) similar to the group label.

Title	Grouping (abstracts)	MeSH keywords
Effects of antenatal dexamethasone treatment on glucocorticoid receptor and calcyon gene expression in the prefrontal cortex of neonatal and adult common marmoset monkeys.	Molecular; Dexamethasone	Age Factors; Animals; Animals, Newborn; **Body Size**; **Body Weight**; Callithrix; Dexamethasone; Female; Glucocorticoids; Male; Membrane Proteins; Prefrontal Cortex; Pregnancy; Prenatal Exposure Delayed Effects; Receptors, Glucocorticoid; Receptors, Mineralocorticoid; RNA, Messenger
The body politic the relationship between stigma and obesity-associated disease.		Adiposity; Age Factors; **Body Mass Index**; Electric Impedance; Female; Humans; Male; **Obesity**; Prejudice; Risk Factors; Sex Factors; Stress, Psychological
Prenatal Stress or High-Fat Diet Increases Susceptibility to Diet-Induced Obesity in Rat Offspring.	High-fat Diet	Animals; Child; Diabetes Mellitus, Type 2; **Dietary Fats**; **Energy Intake**; Female; Genetic Predisposition to Disease; Humans; Infant; Male; **Obesity**; Pregnancy; Prenatal Exposure Delayed Effects; Rats; Rats, Sprague-Dawley
The TNF-System Functional Aspects in Depression, Narcolpdfy and Psychopharmacology.		

a nutshell, this example illustrates our core goal, which is to provide additional, meaningful clusters, which would guarantee high coverage (a small number of unassigned documents left in the result set). We will return to this problem in further analyses and show a trade-off between coverage and specificity of a clustering.

4.5 Clustering Stability

The general framework for comparing different clustering methods that underlies analysis in this section (and two subsequent subsections) is to consider similarity relations induced by these groupings. In other words, we work in the space of pairs of documents and label each pair as "similar" or "dissimilar" according to a given clustering. We interpret documents belonging to "Other topics" group as

dissimilar[2]. If either clustering plays the role of a benchmark or a baseline, the similarity relation induced by this clustering may be considered a decision class, whereas the similarity relation induced by other clustering(s) may be interpreted as a classification model(s).

The goal of a clustering based on a different representation is never to reproduce the baseline grouping, but to provide an alternative one. Nevertheless, we can interpret "accuracy" in this context as a metric of stability of the clustering algorithm with respect to document representation. This metric, when applied to clustering comparison (in the space of pairs of documents) is called Rand Index [8].

We will reiterate two questions already mentioned earlier in this paper:

- How does a clustering change if a baseline representation (based on abstracts) is enriched by including citations?
- How does it change when the representation is replaced with one based on DBpedia concepts?

Graph 3 shows Rand Index calculated against the benchmark clustering. Most noticeable is an overall remarkably high Rand Index for most clusterings (around 0.9), while being significantly lower for clustering based on 300 DBpedia entries "closest" to a given article (we refer to this grouping as to "DBpedia-300" from now on, as this particular clustering will be highly illustrative further in the article).

This analysis shows that extended (or alternative) representations result in overall similar clusterings. The more DBpedia concepts used in document description, the less similar the resulting clustering to the baseline.

A care needs to be taken when interpreting the Rand Index, as this metric does not account for different proportions of similar to dissimilar document pairs (the issue is similar to using accuracy as an assessment of classification with rare classes). While most pairs of documents are dissimilar in most models, "DBpedia-300" induces a much "softer" notion of similarity, thus labeling many more pairs of documents as similar.

4.6 Do Different Document Representations Convey Common Information?

Our next analysis focuses on clustering based solely on bibliographical references. For each document, the set of bibliographical references is disjoint with its' abstract, i.e. bibliographical references correspond to different coordinates in the vector space model than words. We compare similarity matches and mismatches of grouping based on this representation against the benchmark in a contingency table, and calculate Pearson's Chi-squared test with Yates' continuity correction. This procedure yields a p-value 3.42×10^{-06}, remarkably low,

[2] We stress that "Other topics" is an artificial group that consists of articles unassigned to any resulting cluster. One could argue that it conceptually corresponds to a set of singleton clusters.

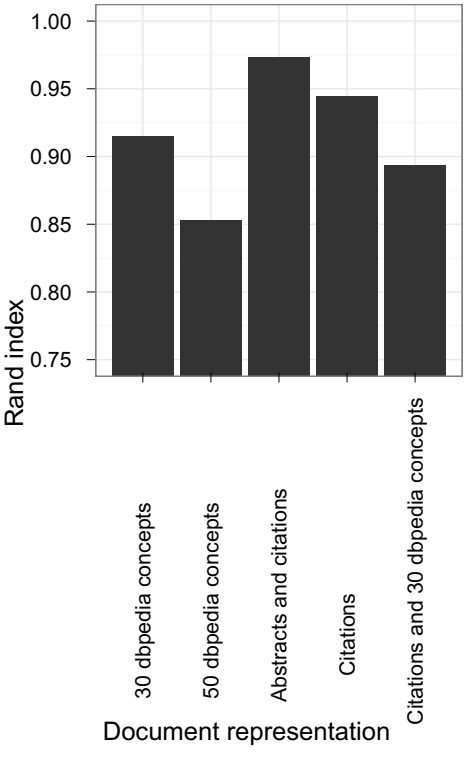

Fig. 3. Rand Index w.r.t. clustering based on abstracts (the benchmark)

despite the general tendency of Yates' continuity correction to underestimate statistical significance. Thus, we conclude that clusterings based on abstracts and those based on references/citations are not independent, with confidence far exceeding 99% ("far" ratio-wise, rather than in absolute terms). However trivial this observation may be, structural content of articles used in our experiment significantly overlaps with their lexical content.

4.7 Validation Using MeSH Vocabulary

Figure 4 shows the results of validation of clusterings using MeSH terms associated with each article. Since these tags were not used in either document representation, they provide a natural confirmatory source of information. Nevertheless, it is important to stress that these terms are not assigned to articles automatically, but by subject analysts.

For each pair of articles we define their distance as the fraction of MeSH terms associated with exactly one of these articles to the overall number of MeSH terms associated with either of these articles. Figure 4 shows the average distance between pairs of documents within the same clusters, with the average taken over all such pairs (rather than over clusters).

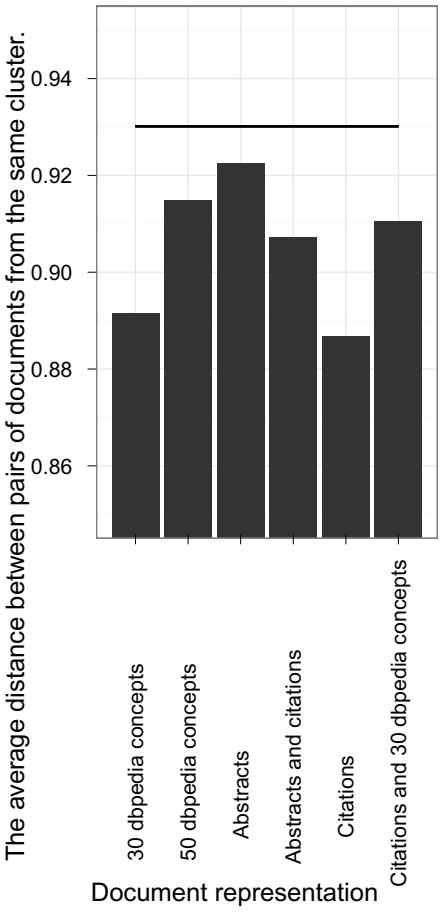

Fig. 4. Validation of clusterings based on different document representations using MeSH terms associated with articles. For each document representation (and hence clustering), we calculate the average distance between documents belonging to the same cluster. The average is taken over all such pairs rather than over all clusters. The vertical line corresponds to the average distance between two documents taken at random.

As the graph suggests, all document representations investigated in our analyses lead to plausible clusterings. However, in order to formulate conclusions of a comparatory nature, we have yet to conduct experiments using different search queries. Moreover, structural information provided with MeSH can be used to define various distance metrics between these terms, and each such a metric can be extended to a distance between documents. Further yet, we plan to use other evaluation metrics, although care needs to be taken, as the clustering algorithm under consideration does not output disjoint clusters, neither are all documents assigned to a cluster.

4.8 Structural Properties of Clusterings

The left plot on figure 5 shows that "DBpedia-300" induces a very vague similarity between documents, as each document is assigned to approximately three different clusters on the average. Remaining clusterings are more conservative in this aspect, although not fully unambiguous either.

Figure 6 shows the trade-off between the "vagueness" of a clustering and the "coverage" (i.e. the number of documents with assigned proper clusters). Clusterings based on 30 (and 50) entries from DBpedia seem to balance this trade-off fairly well. Moreover, they are attractive for a different reason – they provide clusterings that are only indirectly based on document content (e.g. a cluster label may be an entry in DBpedia which is not directly worded in any paper, yet may be highly indicative of a general concept linking a group of articles).

It is worth stressing that a proper balance of these structural properties is crucial for grouping documents in the context of Document Retrieval, while not necessarily in the context of Web Search. When speaking of grouping documents in Web Search, we may restrict our discussion to non-navigational search queries. These queries (due to their vagueness) would usually return a very large number

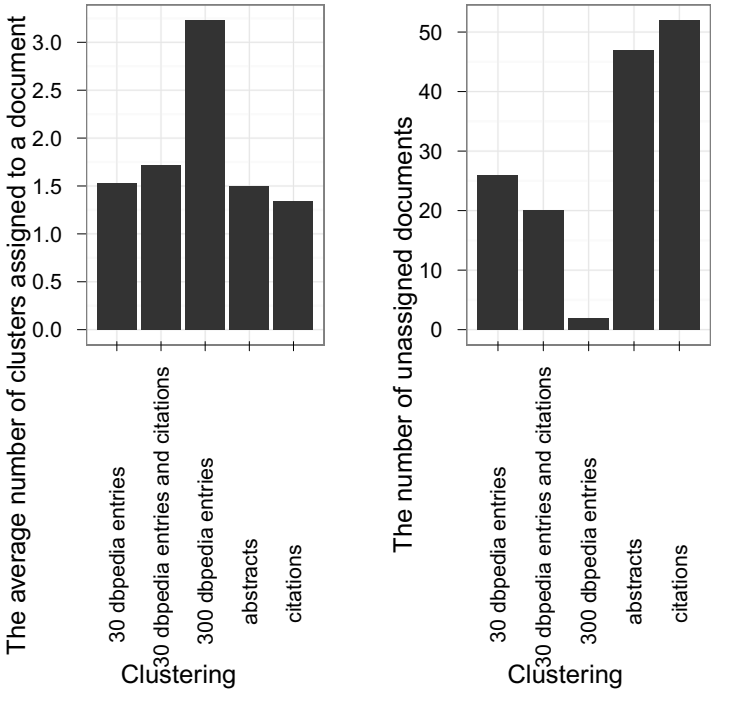

Fig. 5. The average number of clusters assigned to a document (left) and the number of unassigned documents (right) for selected clusterings

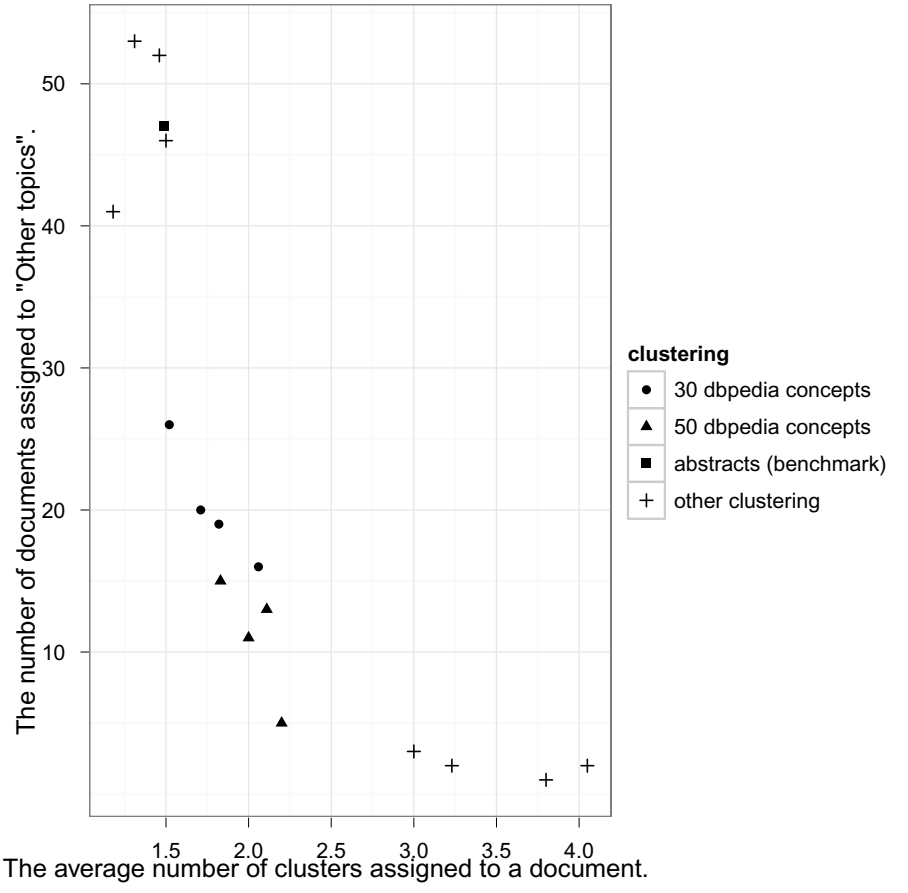

Fig. 6. The trade-off between clustering specificity and coverage. Points to the right correspond to clusterings based on the richest representations – those which include 300 DBpedia concepts.

of matching webpages. A clustering tool usually limits the number of processed snippets to a small subset (e.g. 200) and relies on the interaction with the user to further refine his query or navigate the result set by other means. On the other hand, in Document Retrieval we are always limited to a fixed (often small) number of matching documents, and our intent is to provide plausible grouping which preferably describes the vast majority of them.

5 Conclusions and Further Work

Our preliminary experiments lead to several promising insights, although a wider coverage of experiments need to be conducted in order to speak of definite

conclusions, which would translate to a wide range of queries (or rather, result sets) and text corpora. Nevertheless, our analyses suggest the following:

- Certain documents which would be naturally grouped together do not explicitly share much common lexical content in their abstracts, whereas other document representations convey such information, hence such representations may be used to supplement or provide alternative clusterings.
- The clustering algorithm used in our experiments (LINGO) is stable with respect to the input data. In other words, similarity relations induced by clusterings based on close document representations also yield a high degree of similarity. This suggests that a carefully selected extension method could in fact yield a natural refinement of a baseline clustering.
- Validation using MeSH terms suggests that all document representations investigated in this paper lead to plausible results, although our current analyses focused on clusters without regard to labels. Label evaluation is yet to be conducted.
- Varying richness of document representation displays a trade-off between important structural properties of resulting clusters, namely – the specificity of groups and the number of unassigned documents. Hence, results that confirm closer to a postulated balance may be found if we prepare appropriate document representations.

Our future plans are briefly outlined as follows:

- Use additional information from MeSH vocabulary (e.g. structural information about relationships between terms),
- analyse label quality of clusters resulting from different document representations,
- use MeSH for document representation or label assignment rather than merely for validation of results,
- conduct experiments using other extensions (e.g. citations along with their context; information about authors, institutions, fields of knowledge or time),
- grouping of objects of other types (e.g. authors, institutions, ...),
- visualization of clustering results.

References

1. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval, 1st edn. Addison Wesley Longman Publishing Co. Inc. (May 1999)
2. Carpineto, C., Osiński, S., Romano, G., Weiss, D.: A survey of web clustering engines. ACM Comput. Surv. 41, 17:1–17:38 (2009)
3. Hearst, M.A., Pedersen, J.O.: Reexamining the cluster hypothesis: Scatter/gather on retrieval results. In: Proceedings of SIGIR 1996, 19th ACM International Conference on Research and Development in Information Retrieval, Zürich, CH, pp. 76–84 (1996)
4. Komorowski, J., Pawlak, Z., Polkowski, L., Skowron, A.: Rough sets: a Tutorial (1998)

5. Ngo, C.L., Nguyen, H.S.: A method of web search result clustering based on rough sets. In: Skowron, A., Agrawal, R., Luck, M., Yamaguchi, T., Morizet-Mahoudeaux, P., Liu, J., Zhong, N. (eds.) Web Intelligence, pp. 673–679. IEEE Computer Society (2005)
6. Osinski, S.: An algorithm for clustering of web search result. Master's thesis, Poznan University of Technology, Poland (June 2003)
7. Pawlak, Z.: Rough Sets: Theoretical Aspects of Reasoning About Data. Kluwer, Dordrecht (1991)
8. Rand, W.: Objective criteria for the evaluation of clustering methods. Journal of the American Statistical association 66(336), 846–850 (1971)
9. Roberts, R.J.: PubMed Central: The GenBank of the published literature. Proceedings of the National Academy of Sciences of the United States of America 98(2), 381–382 (2001)
10. Kawasaki, S., Nguyen, N.B., Ho, T.-B.: Hierarchical document clustering based on tolerance rough set model. In: Zighed, D.A., Komorowski, J., Żytkow, J.M. (eds.) PKDD 2000. LNCS (LNAI), vol. 1910, pp. 458–463. Springer, Heidelberg (2000)
11. Skowron, A., Stepaniuk, J.: Tolerance approximation spaces. Fundamenta Informaticae 27(2-3), 245–253 (1996)
12. Szczuka, M., Janusz, A., Herba, K.: Clustering of rough set related documents with use of knowledge from dBpedia. In: Yao, J. (ed.) RSKT 2011. LNCS, vol. 6954, pp. 394–403. Springer, Heidelberg (2011)
13. Ho, T.B., Nguyen, N.B.: Nonhierarchical document clustering based on a tolerance rough set model. International Journal of Intelligent Systems 17(2), 199–212 (2002)
14. Weiss, D.: A clustering interface for web search results in polish and english. Master's thesis, Poznan University of Technology, Poland (June 2001)
15. Wroblewski, M.: A hierarchical www pages clustering algorithm based on the vector space model. Master's thesis, Poznan University of Technology, Poland (July 2003)
16. Zamir, O., Etzioni, O.: Grouper: a dynamic clustering interface to web search results. Computer Networks 31(11–16), 1361–1374 (1999)