RELFIN - Topic Discovery for Ontology Enhancement and Annotation*

Markus Schaal, Roland M. Müller, Marko Brunzel, and Myra Spiliopoulou

Otto-von-Guericke-University Magdeburg
forename.name@iti.cs.uni-magdeburg.de

Abstract. While classic information retrieval methods return whole documents as a result of a query, many information demands would be better satisfied by fine-grain access inside the documents. One way to support this goal is to make the semantics of small document regions explicit, e.g. as XML labels, so that query engines can exploit them. To this purpose, the *topics* of the small document regions must be discovered from the texts; differently from document labelling applications, fine-grain topics cannot be listed in advance for arbitrary collections. Text-understanding approaches can derive the topic of a document region but are less appropriate for the construction of a *small set of topics* that can be used in queries.

To address this challenge we propose the coupling of text mining, prior knowledge explicated in ontologies and human expertise and present the system RELFIN, which is designed to assis the human expert in the *discovery of topics* appropriate for (i) ontology enhancement with additional concepts or relationships, (ii) semantic characterization and tagging of document regions. RELFIN performs *data mining* upon linguistically preprocessed corpora to group document regions on *topics* and *constructing the topic labels* for them, so that the labels are characteristic of the regions and thus helpful in ontology-based search. We show our first results of applying RELFIN on a case study of text analysis and retrieval.

Keywords: Topic Discovery, Label Construction, Ontology Enhancement, Text Clustering.

1 Introduction

Ontologies over document corpora and semantic labels inside the documents can greatly enhance information acquisition: Ontologies describe concepts and the relationships among them and map them into their textual representations in the documents. A semantic label reflects the topic of a small part of a document, e.g. a paragraph or a sentence; if it is implemented as an annotation tag, it can

* Work partially funded under the EU Contract IST-2001-39023 Parmenides. http://www.crim.co.umist.ac.uk/parmenides

A. Gómez-Pérez and J. Euzenat (Eds.): ESWC 2005, LNCS 3532, pp. 608–622, 2005.

be exploited by a query engine. The topics corresponding to the semantic labels may or may not consist of terms from the ontology, so the two instruments support information acquisition in complementing ways.

In this paper, we present the interactive system RELFIN for the discovery of region-level topics in documents. RELFIN is part of the PARMENIDES integrated environment, which encompasses tools for linguistic pre-processing, ontology enhancement through concepts and relations, semantic text annotation and extraction of entities and events. RELFIN uses data mining techniques to analyse and group semantically similar document regions and to derive labels as topics from them. At the same time, RELFIN interacts with the human expert who provides the context knowledge and assists her by proposing topics for text annotation or ontology enhancement.

Central to RELFIN is the notion of *topic cluster*. A topic cluster is a non-marginal set of similar document regions, where similarity is given by the cosine-distance between vectors. Each document region is represented by a vector over a feature space of concepts from an ontology. The weights are computed by the TF-IDF value of the term count for a feature. A cluster label is constructed as a concatenation of features with a high support within the cluster - the so-called *frequent features*. We term the generation of topic clusters as *topic discovery*. Topic discovery can be used both for ontology enhancement and text annotation but it must be stressed, that these tasks are distinct. We propose the use of additional corpus specific terms for ontology enhancement and a novel quality criterium for region annotation.

For *ontology enhancement*, new terms are proposed as new concepts, while groups of co-occuring terms (concepts) are indicatory of the need to create links among them. Thus, the human expert is supported in creating added value by juxtaposing her background knowledge to corpus content and exploiting both to enrich the ontology. It should be noted that even the richest ontology may need this type of enrichment: A corpus focusses on specific aspects of a universe of discourse, which may or may not be explicit in the ontology. Moreover, the individual document regions may refer to topics that turn to be a posteriori of importance for the ontology. However, topics at region level may be too fine-grain for ontology enhancement but nonetheless appropriate for search inside the specific corpus. Hence, RELFIN supports *text annotation*, i.e. the tagging of document regions with the derived topic labels, as complementary task to ontology enhancement.

A major challenge for topic clustering is the specification of the label. Obviously, a label consisting of the single term “be” is not informative for most corpora. A label consisting of 100 terms appearing in some of the cluster members is not very useful towards a search engine either. For RELFIN, we propose a novel criterium for cluster labelling - the *Residuum* of non-frequent features within the topic cluster: First, RELFIN clusters document regions on similarity and derives labels for the clusters. Then, a residuum threshold is set and clusters not satisfying the threshold are rejected. The retained topic clusters with their labels are proposed to the domain expert as *labelled topic clusters*.

In the next section, we discuss related work. In section 3 we introduce the PARMENIDES framework, in which RELFIN operates and then elaborate on RELFIN in section 4. Section 5 contains a first set of experiments for interactive ontology enhancement and text annotation and a discussion of the findings. The last section concludes the study.

2 Related Work

There is increasing research on the discovery of semantic labels from text data. Some methods put their emphasis on the formulation of appropriate labels [MB01, RM99, GSW01, WS01b], while others further consider the establishment of schemata or other semantic descriptions from those labels [HSS03, KVM00, MS00a, MS00b, WS01a, WS02]. We elaborate on these two types of methods in the following. Moore and Berman propose an algorithm that converts textual pathology reports into XML documents: Natural Language Processing (NLP) techniques are applied upon the texts; the identified terms and noun groups are mapped upon concepts of a medical thesaurus; these concepts become XML tags that annotate the corresponding pieces of text [MB01]. This approach can achieve an extensive annotation of the corpus at a great level of detail.

Rauber and Merkl derive document labels by a clustering technique [RM99]: The documents are modelled as vectors of weighted terms and clustered on similarity. For each cluster thus established, a label is derived by considering the terms characterizing the cluster. The core methodology is conceptually the same as in our previous work on the derivement of labels for sentences by similarity-based clustering of sentence contents [GSW01, WS01b].

Handschuh et. al. [HSV03] presents a system that learns information extraction rules from manually tagged input. In contrast to our approach they focus on entities extraction and not on topic discovery and they need pre-labelled documents that we don't need.

The subject of deriving an appropriate semantic label for a set of similar texts is also addressed in [HSS03]: Hotho et al use text clustering to derive text clusters. However they subsequently use formal concept analysis to construct a concept lattice and don't use metrics to check the validity of a cluster for topic enhancement.

The extraction of a domain-specific ontology from texts with data mining techniques is discussed in [KVM00, MS00a, MS00b], whereby [KVM00, MS00b] concentrate on the core mechanism, which relies on the frequency of concepts in the texts, while the emphasis in [MS00a] is on the discovery of semantic relations by using association rules. The semantic richness and diversity of corpora does not lend itself to full automation, so that the involvement of a domain expert becomes necessary [MS00c].

The ASIUM system [FN99] uses unsupervised concept clustering methods to learn semi-automatically subcategorization frames of verbs and ontologies. However they haven't used text unit clustering and don't use cluster quality criteria.

In this study, we extend our previous work on the "DIAsDEM Workbench" for the formulation of semantic labels for text fragments [GSW01, WS02]. Similarly to the original DIAsDEM Workbench, we perform clustering over the document corpus to establish a set of clusters, for which semantic labels can be derived. However, we replace the original rudimentary criteria on cluster cardinality and number of representative features in a cluster with a more sophisticated measure of cluster quality, the so-called *residuum* of non-frequent features, thereby enabling the automatic selection and labelling of high-quality clusters.

3 Parmenides Framework

PARMENIDES is an EU-funded project in the area of knowledge extraction and management. One of its goals is the realization of an ontology-driven systematic approach for integrating the entire process of information gathering, processing and analysis (cf. [SRB$^+$04]).

One task within this goal is the extraction of knowledge from texts. Knowledge extraction is directed towards (a) the establishment of ontologies which reflect the universe of discourse and (b) the semantic annotation of documents with the concepts, entities and events depicted in the ontologies.

The RELFIN module is responsible for ontology enhancement with new concepts and with concept groups, as well as the semantic annotation of texts with such concepts/groups. As explained in the introduction, we use the collective term "topic" for them. This knowledge extraction process involves at least one human expert, who aligns knowledge extraction to the business objectives by:

- providing an initial ontology
- enhancing the ontology with concepts and relations found by the "RELFIN Learner"
- reviewing the topic clusters and proposed labels for the annotation to be performed by the "RELFIN Annotator"

The software components RELFIN Learner and RELFIN Annotator can be seamlessly integrated into "PARMENIDES workflows": A workflow is a series of component invocations that can be specified graphically and then executed on the fly as shown in Fig. 1. The components process documents and enrich them with annotations at different levels of complexity and semantic [RDH$^+$03a], i.e. by representing linguistic as well as conceptual knowledge. The XML-based ParDoc format [RDH$^+$03b] is used as reference format. The components interact with each other via a document queue, depicted in the figure under the label `NormalQueue`.

The example workflow of Fig. 1 consists of the following components[1] (named by their labels):

[1] Printed with the kind agreement of the responsible PARMENIDES partners

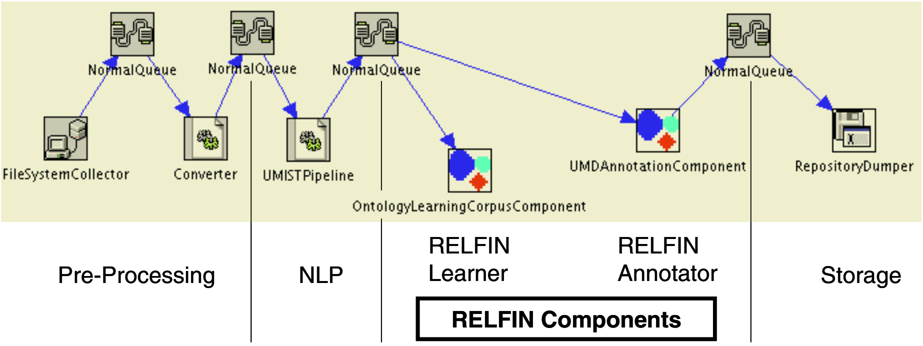


Fig. 1. An example PARMENIDES workflow (screenshot + annotation)

`FileSystemCollector`: This component collects documents from a repository on a hard disk. For collecting documents from the web, the component WebSystemCollector should be used instead.

`Converter`: This component is responsible for creating ParDoc documents from other document types, such as HTML, plain text, pdf, word and ppt.

`UMISTPipeline`: This component is a four-step analysis process that performs basic NLP and information extraction functionality. Its functionality consists of Tokenization, POS Tagging, Sentence Splitting, Ontology Lookup operations and ontology-based Information Extraction using the Cafetiere software [2].

`OntologyLearningCorpusComponent`: This component is the RELFIN Learner component for text clustering and interactive expert involvement, as described in section 4. It takes as input a whole collection of documents (a seed collection) and outputs clusterings, expert-approved topic labels and the documents with annotations.

`UMDAnnotationComponent`: This component is the RELFIN Annotator component. It reads as input a document, as well as the clusterings and the approved cluster labels output by the RELFIN Learner. It assigns the regions of the document into clusters and annotates them with the corresponding topic labels.

`RepositoryDumper`: This component stores ParDoc documents into a Document Repository.

4 The RELFIN Learner

In Section 4.1, we give a formal introduction to the concepts used with the RELFIN Learner including the novel cluster quality criterium *Residuum*. We

[2] For a documentation on the UMIST Pipeline, cf. Vasilakopoulos et al. [VBB04]

show the procedure of the RELFIN Learner in Section 4.2 and present some details about the human expert interaction with the RELFIN GUI in Section 4.3.

4.1 Formal Concepts

A *text unit* is an arbitrary text fragment produced by a linguistic tool, e.g. by a sentence-splitter. Text units consist of words. A *term* is a stemmed or lemmatized word. Thus, text units can be measured according to the frequency of the terms contained in it. Text units correspond to documents or document regions, e.g. paragraphs or sentences. A *text corpus* $\mathcal{A} = \{1, \ldots, n\}$ is a set of text units.

A term is a textual representation of a *concept*. Generally, there is a m-to-n mapping between terms in a text corpus and a set of concepts that describe a universe of discourse due to synonyms and homonyms. Terms, concepts and the mapping between them are part of the *ontology* of this universe.

A *feature space* $\mathcal{F} = (1, \ldots, d)$ is a sequence of features, where each feature corresponds to a single concept. A *vectorization* $\mathcal{X}$ of the text corpus is obtained by counting all terms that are mapped to each of the features in the feature space for all text units of a text corpus. Subsequently, TF-IDF weighting is applied[3]. $\mathcal{X}$ is a 2-dimensional matrix given by values x_{ij} per text unit $1 \leq i \leq n$ and feature $1 \leq j \leq d$. Thus, each text unit i is represented by its vector $x_i = (x_{i1}, \ldots, x_{id})$ over the feature space.

A *cluster* $C \subseteq \mathcal{X}$ is a set of vectors. A *cluster label* is a term or term combination that is given by the *frequent features* of a cluster. A *frequent feature* is a feature whose in-cluster support is above a certain threshold τ_{ics}.

Definition 1 (In-Cluster Support of a Feature). *Let $C \subseteq \mathcal{X}$ be a cluster, where $\mathcal{X}$ is the vector space over the text corpus $\mathcal{A}$ for the feature space $\mathcal{F}$. Let $k \in \mathcal{F}$ denote a feature. The* in-cluster support *of feature k in C is the count of vectors $x \in C$ that contain feature k (i.e. $x_k \neq 0$) divided by the cardinality of C.*

$$ics(k, C) = \frac{\mid \{x \in C \mid x_k \neq 0\} \mid}{\mid C \mid} \tag{1}$$

One criteria for clusters having a good label is newly introduced here, the so-called *Residue* of the in-cluster support of infrequent labels. *Topic Clusters* with a residue lower than a certain residue threshold τ_{res} are *pure.*

Definition 2 (Residue). *Let $C \subseteq \mathcal{X}$ be a cluster and let τ_{ics} be the lower boundary to the in-cluster support of features, thus determining which features are frequent. Then, the* residue *of C subject to this threshold is the relative in-cluster support for infrequent features:*

$$residue(C, \tau_{\text{ics}}) = \frac{\sum_{k \in nonfreq(C, \tau_{\text{ics}})} ics(k, C)}{\sum_{k \in \mathcal{F}} ics(k, C)} \tag{2}$$

where $nonfreq(C, \tau_{\text{ics}}) = \{k \in \mathcal{F} \mid ics(k, C) \leq \tau_{\text{ics}}\}$.

[3] For a documentation on Vectorization and TF-IDF weighting, cf. Salton and Buckley [SB88].

4.2 Procedure

The RELFIN procedure is described by a Data Flow Diagram (DFD) in Fig. 2. *Processes* are represented by circles, *external input* is represented by squares and *data stores* are represented by open boxes (over- and underlined). The software proceeds as follows (cf. Fig. 2):

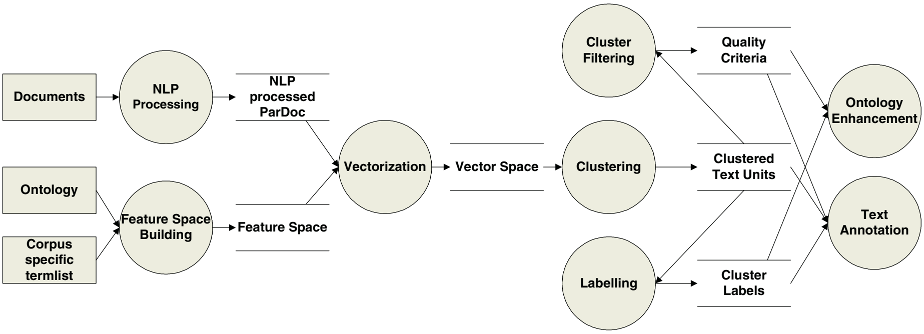


Fig. 2. Data Flow Diagram of RELFIN Learner

NLP Processing. The RELFIN application relies on a NLP processed document collection. These documents are provided in the ParDoc format.

Feature Space Building. The feature space is build from an ontology and/or a list of corpus specific terms[4]. When using the ontology[5], each class including its synonyms and its instances[6] (and their synonyms) becomes a feature of the feature space.

Vectorization. Different text granularities are available from the ParDoc format for vectorization: (1) documents as a whole, (2) paragraphs or (3) sentences. Here we use sentences as the chosen granularity. Each text unit is represented by its vector computed from the feature space. Only text units with two or more non-zero values are used for further processing.

Clustering. The text units are clustered by use of a Bi-Secting k-Means algorithm [SKK00], which partitions the instances in k clusters. The parameter k is specified by the user. The Bi-Secting k-Means is a variation of the k-means and showed great success in the text clustering problem[SKK00]. The algorithm starts with a single cluster which is split into two clusters by a k-means algorithm with $k = 2$. Then, the biggest cluster is choosen and it is again split in two clusters. This is done until the desired cluster number

[4] Corpus specific terms are ordered by their rank position ratio with respect to a general language corpus, here the British National Corpus.

[5] The PARMENIDES project incorporates an ontology editor for the maintenance of ontologies.

[6] In the used ontology format, instances are maintained as special concepts to be subsumed together with their class concepts.

is reached. Alternatively, the cluster with highest residuum (cf. paragraph *Cluster Filtering* below) is chosen instead of the biggest cluster. The cosine metric is used as the distance function.

Cluster Filtering. After the clustering, the quality measures of the clusters are calculated. For a cluster to be considered as a *labelled topic cluster* (and thus be accepted), we require the cluster

- to be *non-marginal*, i.e. to have a cardinality above a certain threshold c_{min} and
- to be *pure*, i.e. to have a residuum lower than a given threshold τ_{res} (with respect to frequent feature threshold τ_{ics}, cf. Section 4.1).

The label of a *pure* and *non-marginal* cluster is given by the set of its frequent features, i.e. features with high in-cluster support, whereas there are only few instances not covered by the frequent features.

Labelling. For each cluster, the set of frequent features is concatenated and proposed as the cluster label.

Ontology Enhancement. RELFIN can be used for ontology enhancement, cf. Section 4.3.

Text Annotation. RELFIN can be used for semi-automatic annotation, cf. Section 4.3.

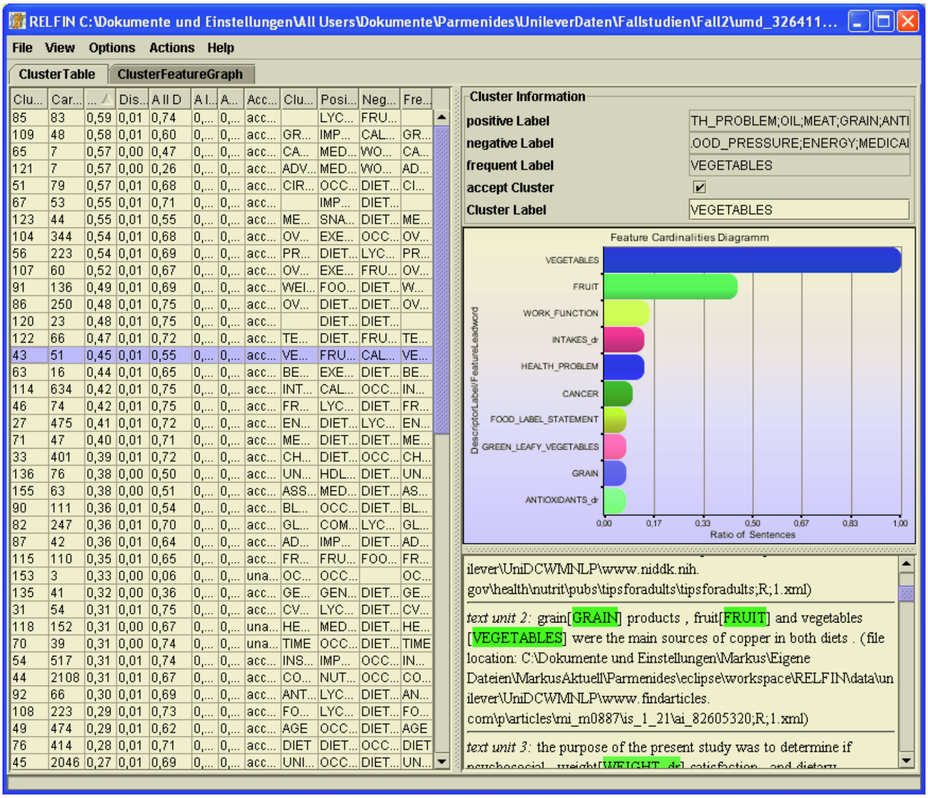
4.3 Human Expert Interaction

Figure 3 shows the RELFIN GUI, displaying a table of clusters on the left side and details of the selected cluster on the right side. The table allows sorting the clusters on certain attributes, associated with the clusters. The bar chart diagram shows the percentage of instances which have a certain feature, for the ten most frequent features of the selected cluster. In the lower right corner examples of text units in the current cluster are displayed, whereas terms, included in the feature space are highlighted. These example text units help the user to justify the appropriateness of a cluster label.

For ontology enhancement, an ontology editor is opened on the desktop next to the RELFIN-GUI, where the domain expert can edit an existing ontology of his choice. Good candidates for enhancement are homogeneous[7] clusters with respect to the cosine distance, it is on experts choice whether to include a certain feature - as a new (1) class (concept) or (2) instance (concept), as a (3) synonym of an existing concept or as a (4) attribute type/ attribute value.

For annotation, a domain expert is required to browse (at least) the accepted clusters in order to deny acceptance by deselecting the check mark next to `accept Cluster` and to optionally edit the `Cluster Label` in the cluster information section of the RELFIN GUI, cf. upper-right corner of Fig. 3.

[7] The term *homogeneous* refers to the cluster criteria *average instance to centroid distance (AICD)* and *average instance-to-instance distance (AIID)*, which are based on the cosine distance used for clustering and also shown in the cluster table.



RELFIN C:\Dokumente und Einstellungen\All Users\Dokumente\Parmenides\UnileverDaten\Fallstudien\Fall2\umd_326411...

Clu...	Car...	...	Dis...	A ll D	A l...	A...	Acc...	Clu...	Posi...	Neg...	Fre...
85	83	0,59	0,01	0,74	0,...	0,...	acc...		LYC...	FRU...	
109	48	0,58	0,01	0,60	0,...	0,...	acc...	GR...	IMP...	CAL...	GR...
65	7	0,57	0,00	0,47	0,...	0,...	acc...	CA...	MED...	WO...	CA...
121	7	0,57	0,00	0,26	0,...	0,...	acc...	ADV...	MED...	WO...	AD...
51	79	0,57	0,01	0,68	0,...	0,...	acc...	CIR...	OCC...	DIET...	CI...
67	53	0,55	0,01	0,71	0,...	0,...	acc...		IMP...	DIET...	
123	44	0,55	0,01	0,55	0,...	0,...	acc...	ME...	SNA...	DIET...	ME...
104	344	0,54	0,01	0,68	0,...	0,...	acc...	OV...	EXE...	OCC...	OV...
56	223	0,54	0,01	0,69	0,...	0,...	acc...	PR...	DIET...	LYC...	PR...
107	60	0,52	0,01	0,67	0,...	0,...	acc...	OV...	EXE...	FRU...	OV...
91	136	0,49	0,01	0,69	0,...	0,...	acc...	WEI...	FOO...	DIET...	W...
86	250	0,48	0,01	0,75	0,...	0,...	acc...	OV...	DIET...	DIET...	OV...
120	23	0,48	0,01	0,75	0,...	0,...	acc...		DIET...	DIET...	
122	66	0,47	0,01	0,72	0,...	0,...	acc...	TE...	DIET...	FRU...	TE...
43	51	0,45	0,01	0,55	0,...	0,...	acc...	VE...	FRU...	CAL...	VE...
63	16	0,44	0,01	0,65	0,...	0,...	acc...	BE...	EXE...	DIET...	BE...
114	634	0,42	0,01	0,75	0,...	0,...	acc...	INT...	CAL...	OCC...	IN...
46	74	0,42	0,01	0,75	0,...	0,...	acc...	FR...	LYC...	DIET...	FR...
27	475	0,41	0,01	0,72	0,...	0,...	acc...	EN...	DIET...	LYC...	EN...
71	47	0,40	0,01	0,71	0,...	0,...	acc...	ME...	DIET...	DIET...	ME...
33	401	0,39	0,01	0,72	0,...	0,...	acc...	CH...	DIET...	OCC...	CH...
136	76	0,38	0,00	0,50	0,...	0,...	acc...	UN...	HDL...	DIET...	UN...
155	63	0,38	0,00	0,51	0,...	0,...	acc...	ASS...	MED...	DIET...	AS...
90	111	0,36	0,01	0,54	0,...	0,...	acc...	BL...	OCC...	DIET...	BL...
82	247	0,36	0,01	0,70	0,...	0,...	acc...	GL...	COM...	LYC...	GL...
87	42	0,36	0,01	0,64	0,...	0,...	acc...	AD...	IMP...	DIET...	AD...
115	110	0,35	0,01	0,65	0,...	0,...	acc...	FR...	FRU...	FOO...	FR...
153	3	0,33	0,00	0,06	0,...	0,...	una...	OC...	OCC...		OC...
135	41	0,32	0,00	0,36	0,...	0,...	acc...	GE...	GEN...	DIET...	GE...
31	54	0,31	0,01	0,75	0,...	0,...	acc...	CV...	LYC...	DIET...	CV...
118	152	0,31	0,00	0,67	0,...	0,...	una...	HE...	MED...	DIET...	HE...
70	39	0,31	0,00	0,74	0,...	0,...	una...	TIME	OCC...	DIET...	TIME
54	517	0,31	0,01	0,74	0,...	0,...	acc...	INS...	IMP...	OCC...	IN...
44	2108	0,31	0,01	0,67	0,...	0,...	acc...	CO...	NUT...	OCC...	CO...
92	66	0,30	0,01	0,69	0,...	0,...	acc...	ANT...	LYC...	DIET...	AN...
108	223	0,29	0,01	0,73	0,...	0,...	acc...	FO...	LYC...	DIET...	FO...
49	474	0,29	0,01	0,62	0,...	0,...	acc...	AGE	OCC...	DIET...	AGE
76	414	0,28	0,01	0,71	0,...	0,...	acc...	DIET	DIET...	OCC...	DIET
45	2046	0,27	0,01	0,69	0,...	0,...	acc...	UNI...	OCC...	DIET...	UN...

Cluster Information

positive Label	TH_PROBLEM;OIL;MEAT;GRAIN;ANTI
negative Label	.OOD_PRESSURE;ENERGY;MEDICAI
frequent Label	VEGETABLES
accept Cluster	☑
Cluster Label	VEGETABLES

ilever\UniDCWMNLP\www.niddk.nih.
gov\health\nutrit\pubs\tipsforadults\tipsforadults;R;1.xml)

text unit 2: grain[GRAIN] products , fruit[FRUIT] and vegetables [VEGETABLES] were the main sources of copper in both diets . (file location: C:\Dokumente und Einstellungen\Markus\Eigene Dateien\MarkusAktuell\Parmenides\eclipse\workspace\RELFIN\data\unilever\UniDCWMNLP\www.findarticles.com\p\articles\mi_m0887\is_1_21\ai_82605320;R;1.xml)

text unit 3: the purpose of the present study was to determine if psychosocial , weight[WEIGHT_dr] satisfaction , and dietary

Fig. 3. Presentation of the clustering results in RELFIN

5 Experiments

For our experiments, we employed a text corpus collected from the internet for the weight management domain. This text corpus and an initial ontology were provided by an Unilever domain expert. An objective of the weight management case study of Unilever is the improvement of information retrieval and decision support. The corpus contains 1394 documents and has been split into more than 20.000 sentence-level text units. For the experiments, not all but only text units with a certain feature support were used, cf. Section 4. We performed two investigations on the text corpus:

1. Topic Clusters generated by clustering and their frequent features were presented to a human expert. The expert was asked to report on the benefits for ontology enhancement.
2. Labelled Topic Clusters were automatically accepted according to their size and residuum. We compared different clustering parameters with respect to their coverage of the text corpus.

The first investigation examined the usability of text clustering for *ontology enhancement*, the second investigation examined the ability of text clustering to find *labelled topic clusters*. Note, that the underlying corpus does not have a gold-standard ontology or annotation.

5.1 Ontology Enhancement

For the purpose of human expert ontology enhancement, 80 clusters were generated. For building the feature space, the features from the ontology were complemented with terms from the corpus-specific term list, so that 500 features were used altogether. The ic-support threshold for frequent features was set to $\tau_{\text{ics}} = 0.2$ and all clusters were presented to the human expert. By creating 80 clusters only, we got good topic clusters for ontology enhancement, but the topic clusters weren't pure enough for annotation.

In order to evaluate the use of proposed clusters for ontology enhancement, the Unilever domain expert was asked to evaluate the topic clusters according to the following criteria:

- Do the given term or term combinations (the frequent features) make sense and is it of relevance in the use case? Please indicate by Accepting/ Rejecting each cluster.
- Are some of the delivered terms or term combinations appropriate for ontology enhancement? A term or term combination is appropriate if you would decide to put it in the ontology. A combination can be put into the ontology as a combined concept or by establishing a link between concepts.

```
------------------------------------
Cluster 27 - Accepted:
Frequent Terms: FAT;ENERGY
Ontology Enhancement:
Link: FAT "is an" ENERGY "source"
Link: FAT is a "component of" a FOOD_PRODUCT
Link: FOOD_PRODUCT "delivers" ENERGY (Joule)
------------------------------------
Cluster 29 - Rejected:
Frequent Terms: DIETARY_cs
------------------------------------
Cluster 31 - Accepted:
Frequent Terms: CVD_cs;HEALTH_PROBLEM
Ontology Enhancement:
Add: CVD
Link: CVD (acronym of cardiovascular disease) "is a" HEART_DISEASE
Link: HEART_DISEASE "is a" HEALTH_PROBLEM
------------------------------------
```

Fig. 4. Expert Contribution for Ontology Enhancement (for 3 sample clusters)

Note that the review of unlabelled topic clusters is not of use with such a low ic-support threshold. Moreover, features below the threshold are of no interest. Therefore, the expert was provided with a report on the frequent labels per cluster only, without asking him to browse the cluster table.

The whole clustering was evaluated by the Unilever domain expert according to the above criteria, a sample of the results is shown in Fig. 4. Frequent features with suffix "_cs" are the ones originating from the corpus specific term list.

Out of 79 clusters, the human expert accepted 30 clusters. Based on the frequent feature combinations of the accepted clusters, he proposed 21 new concepts, 14 new synonyms and 10 new links between concepts for ontology enhancement.

5.2 Labelled Topic Clusters

By our approach of filtering topic clusters for purity by setting a threshold on the residuum, we have deliberately surrendered a full coverage of all text units. Here we study the coverage of text units by labelled topic clusters for different splitting criteria, residuum thresholds, cluster counts and different feature spaces.

In a first experiment, only the initial ontology was used for building the feature space, resulting in 12990 text units to be selected for clustering. Fig. 5 shows the text units covered by topic clusters over the amount of generated clusters. Two different splitting criteria for the Bi-Secting k-Means algorithm have been used, namely *Splitting the Cluster with Highest Cardinality* (Card-Split) and *Splitting the Cluster with the Highest Residuum* (Res-Split). Clusters have been accepted as topic clusters with residuum threshold $\tau_{\text{res}} = 0.5$ (threshold=0.5) and $\tau_{\text{res}} = 0.3$ (threshold=0.3) respectively. In all cases, the minimum cardinal-

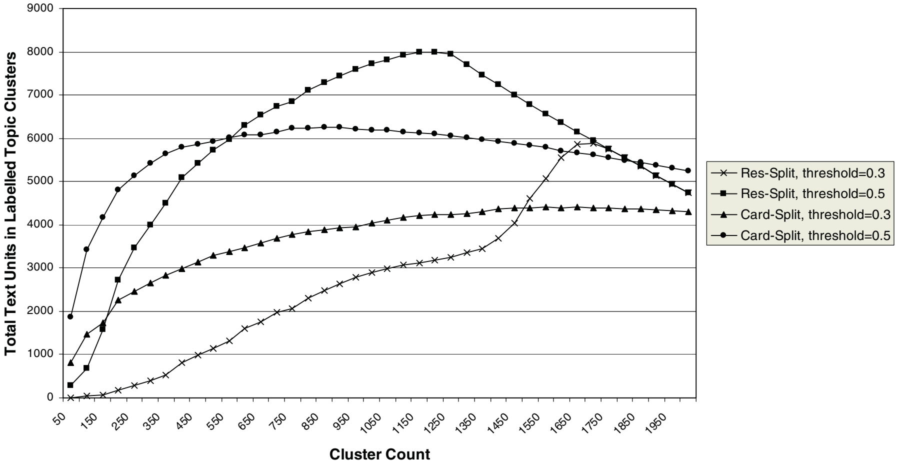


Fig. 5. Text Units in Topic Clusters (total 12990)

ity threshold was $c_{\text{min}} = 15$ and the ic-support threshold for frequent features was $\tau_{\text{ics}} = 0.8$.

For the case of *Splitting the Cluster with the Highest Residuum* (Res-Split), the maximum is reached quite late, i.e. 8.000 text units at 1.150 clusters (147 topic clusters) for the case of $\tau_{\text{res}} = 0.5$ and 5.880 text units at 1700 clusters (133 topic clusters) for the case of $\tau_{\text{res}} = 0.3$.

For the case of *Splitting the Cluster with Highest Cardinality* (Card-Split), the maximum is considerably lower (most likely due to the creation of marginal clusters), but better residue can be reached with less clusters.

In a second experiment, we compared the use of an initial ontology with the use of corpus specific terms. Fig. 6 shows the result for using 300 terms of an ontology (Ontology) juxtaposed against using the first 300 corpus specific and using both (600 words). All computations have been performed with *Res-Split* and residuum threshold $\tau_{\text{res}} = 0.5$ (threshold=0.5). Note that the total size of text units for the clustering varies, since different feature spaces are build and only vectors with two or more non-zero values are accepted.

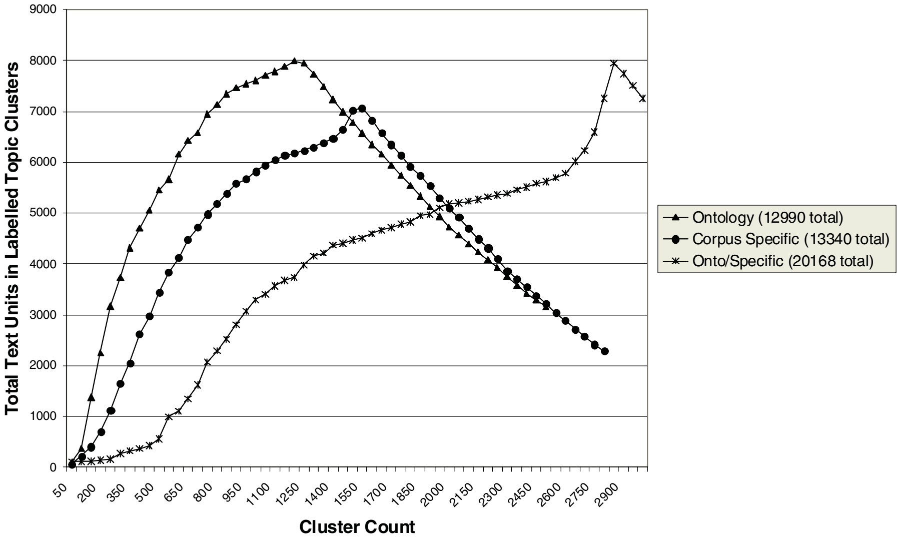


Fig. 6. Text Units in Topic Clusters (Res-Split, threshold=0.5)

For the case of the ontology, the maximum is higher than with the corpus specific term-list, while the maximum of the combination of both is reached later. It should be noted, that the usage of 600 words in the combined case allowed for a total of 20.168 text units to be represented by a vector with two or more non-zero values. The word-list's weakness when compared with the ontology (while actually processing more text units, 13.340 vs. 12.990) might be caused by the ontology being properly tailored towards the text corpus by the domain expert.

6 Conclusions

We presented the fully-fledged RELFIN application as an integrated component within the PARMENIDES framework and showed its ability to support semi-automatic ontology enhancement and a novel approach of filtering for pure topic clusters.

We applied the RELFIN methodology to a real use case without pre-existing gold standards for ontologies or text annotation and learned from first experiments:

- The stimulation of the human expert by looking at topic clusters is manifold and leads to added value by knowledge explication. The use of an ontology editor in parallel to the RELFIN software is suggested.
- The two-phase approach for discovering labelled topics may reach a good coverage of the text corpus, at least with a domain-specific ontology that is tailored towards supporting annotation.

In the future, we intend to further improve the integration of technologies and expert interaction models for semi-automatic ontology enhancement and annotation - possibly with focus on semantic web evolution.

Acknowledgments: We would like to thank the Parmenides consortium and especially the partner Unilever for their contribution to the experiment.

References

[FN99] David Faure and Claire Nédellec. Knowledge acquisition of predicate argument structures from technical texts using machine learning: the system ASIUM. In Dieter Fensel and Rudi Studer, editors, *Knowledge Acquisition, Modeling and Management: 11th European Workshop, EKAW '99, Dagstuhl Castle, Germany, May 1999: Proceedings*, volume 1621 of *Lecture Notes in Computer Science*, pages 329–334. Springer-Verlag, Heidelberg, 1999.

[GSW01] Henner Graubitz, Myra Spiliopoulou, and Karsten Winkler. The DIAsDEM framework for converting domain-specific texts into XML documents with data mining techniques. In *Proc. of the 1st IEEE Intl. Conf. on Data Mining,*, pages 171–178, San Jose, CA, Nov. 2001. IEEE.

[HSS03] Andreas Hotho, Steffen Staab, and Gerd Stumme. Explaining text clustering results using semantic structures. In *Proc. of ECML/PKDD 2003*, LNAI 2838, pages 217–228, Cavtat-Dubrovnik, Croatia, Sept. 2003. Springer Verlag.

[HSV03] Siegfried Handschuh, Steffen Staab, and Raphael Volz. On deep annotation. In *Proceedings of the Twelfth International Conference on World Wide Web*, pages 431–438, Budapest, Hungary, May 2003. ACM Press.

[KVM00] Jörg-Uwe Kietz, Raphael Volz, and Alexander Maedche. Extracting a domain-specific ontology from a corporate intranet. In Claire Cardie, Walter Daelemans, Claire Nédellec, and Erik Tjong Kim Sang, editors, *Proc. of 4th Conf. on Computational Natural Language Learning and of the 2nd Learning Language in Logic Workshop*, pages 167–175, Somerset, New Jersey, 2000. Association for Computational Linguistics.

[MB01] G. William Moore and Jules J. Berman. Medical data mining and knowledge discovery. In *Anatomic Pathology Data Mining*, volume 60 of *Studies in Fuzziness and Soft Computing*, pages 72–117, Heidelberg, New York, 2001. Physica-Verlag.

[MS00a] Alexander Maedche and Steffen Staab. Discovering conceptual relations from text. In *Proc. of ECAI'2000*, pages 321–325, 2000.

[MS00b] Alexander Maedche and Steffen Staab. Mining ontologies from text. In *Proc. of Knowledge Engineering and Knowledge Management (EKAW 2000)*, LNAI 1937. Springer, 2000.

[MS00c] Alexander Maedche and Steffen Staab. Semi-automatic engineering of ontologies from text. In *Proc. of 12th Int. Conf. on Software and Knowledge Engineering*, Chicago, IL, 2000.

[RDH$^+$03a] F. Rinaldi, J. Dowdall, M. Hess, J. Ellman, G. P. Zarri, A. Persidis, L. Bernard, and H. Karanikas. Multilayer annotations in parmenides. In *Proceedings of the K-CAP2003 workshop on "Knowledge Markup and Semantic Annotation"*, October 2003.

[RDH$^+$03b] Fabi Rinaldi, James Dowdall, Michael Hess, Kaarel Kaljurand, Andreas Persidis, Babis Theodoulidis, Bill Black, John McNaught, Haralampos Karanikas, Argyris Vasilakopoulos, Kelly Zervanou, Luc Bernard, Gian Piero Zarri, Hilbert Bruins Slot, Chris van der Touw, Margaret Daniel-King, Nancy Underwood, Agnes Lisowska, Lonneke van der Plas, Veronique Sauron, Myra Spiliopoulou, Marko Brunzel, Jeremy Ellman, Giorgos Orphanos, Thomas Mavroudakis, and Spiros Taraviras. Parmenides: an opportunity for ISO TC37 SC4? In *ACL-2003 workshop on Linguistic Annotation*, Sapporo, Japan, July 2003.

[RM99] Andreas Rauber and Dieter Merkl. Mining text archives: Creating readable maps to structure and describe document collections. In *Principles of Data Mining and Knowledge Discovery*, pages 524–529, 1999.

[SB88] Gerard Salton and Chris Buckley. Term weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523, 1988.

[SKK00] M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. In *KDD Workshop on Text Mining*, 2000.

[SRB$^+$04] Myra Spiliopoulou, Fabio Rinaldi, William J. Black, Gian Piero Zarri, Roland M. Mueller, Marko Brunzel, Babis Theodoulidis, Giorgos Orphanos, Michael Hess, James Dowdall, John McNaught, Maghi King, Andreas Persidis, and Luc Bernard. Coupling information extraction and data mining for ontology learning in parmenides. In *RIAO'2004, April 26th-28th*, Avignon, 2004.

[VBB04] A. Vasilakopoulos, M. Bersani, and W.J. Black. A suite of tools for marking up textual data for temporal text mining scenarios. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC-2004)*, Lisbon, 2004.

[WS01a] Karsten Winkler and Myra Spiliopoulou. Extraction of semantic XML DTDs from texts using data mining techniques. In *Proceedings of the K-CAP 2001 Workshop on Knowledge Markup and Semantic Annotation*, pages 59–68, Victoria, BC, Canada, October 2001.

[WS01b] Karsten Winkler and Myra Spiliopoulou. Semi-automated XML tagging of public text archives: A case study. In *Proceedings of EuroWeb 2001 "The Web in Public Administration"*, pages 271–285, Pisa, Italy, December 2001.

[WS02] Karsten Winkler and Myra Spiliopoulou. Structuring domain-specific text archives by deriving a probabilistic XML DTD. In *6th European Conf. on Principles and Practice of Knowledge Discovery in Databases, PKDD'02*, pages 461–474, Helsinki, Finland, Aug. 2002. Springer Verlag.