Knowledge-Based Systems 24 (2011) 1261–1276

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

OntoPlus: Text-driven ontology extension using ontology content, structure and co-occurrence information

Inna Novalija *, Dunja Mladenić, Luka Bradeško

Artificial Intelligence Laboratory, Jozef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia

ARTICLE INFO

Article history:
Received 1 July 2010
Received in revised form 20 April 2011
Accepted 1 June 2011
Available online 12 June 2011

Keywords:
Knowledge engineering methodologies
Ontology extension
Large-scale ontology
Text mining
Semantic technologies

ABSTRACT

This paper addresses the process of semi-automatic text-driven ontology extension using ontology content, structure and co-occurrence information. A novel **OntoPlus** methodology is proposed for semi-automatic ontology extension based on text mining methods. It allows for the effective extension of the large ontologies, providing a ranked list of potentially relevant concepts and relationships given a new concept (e.g., glossary term) to be inserted in the ontology. A number of experiments are conducted, evaluating measures for ranking correspondence between existing ontology concepts and new domain concepts suggested for the ontology extension. Measures for ranking are based on incorporating ontology content, structure and co-occurrence information. The experiments are performed using a well known Cyc ontology and textual material from two domains – finances and, fisheries & aquaculture. Our experiments show that the best results are achieved by combining content, structure and co-occurrence information. Furthermore, ontology content and structure seem to be more important than co-occurrence for our data in the financial domain. At the same time, ontology content and co-occurrence seem to have higher importance for our fisheries & aquaculture domain.

1. Introduction

This paper explores the process of ontology extension motivated by potential usage of the extended ontology for the analysis of textual information.

For instance, in question answering based on articles about fishery using semantic information from the domain enables providing better answers [1], especially if the semantic information matches content of the articles. Using ontologies allows to search not only within the terms occurring in the query, but also within their semantically related concepts. Given the query "What distinguishes fish?", ontology based system, which performs a search on the fishery related articles, would provide a user with an answer "Various molecular *markers* have been utilized to *distinguish* among easily misidentified *sharks*". Therefore, the available ontology information about *sharks* as a subclass of *fish* leads to more efficient textual data analysis.

Gruber [2] defined **Ontology** as an explicit specification of a conceptualization. According to Gruber [2], ontologies consist of the following main components: concepts, relations, functions, axioms and instances. Ontologies enable effective domain knowledge representation, knowledge sharing and knowledge reuse [3]. Ontologies have been used for different tasks including web page annotation and information retrieval [4], question answering [1,5], word sense disambiguation [6]. The ontology-driven hypothesis generation was successfully applied by Moss et al. [7] in medical domain. Sánchez et al. [8] presented a way of computing concept information content from the Web using ontologies.

The importance of the ontology extension is interconnected with the dynamic nature of the ontologies. When extending large ontologies with new concepts, it is necessary to identify their equivalent concepts already present in the ontology. It is also important to find the correct location and context for new concepts we insert into the ontology. In this paper we present **OntoPlus** methodology, which facilitates complicated, time-consuming and expensive manual development of a large ontology, such as Cyc Knowledge Base (Cyc KB) [9]. Cyc Knowledge Base is a common sense ontology, which is being developed for more than 20 years (more than 900 human years of effort) and is used as a knowledge source in Cyc Artificial Intelligence system. It aggregates already more than 15,000 predicates, 300,000 concepts and 3,500,000 assertions, but the knowledge is still very sparse in various domains. For example, the annotation experiments, conducted on a randomly selected subset of business news, have shown the insufficient representation of financial domain in Cyc and the ways to effectively improve it by the means of Cyc Knowledge Base extension [10]. Cyc is characterized as relatively sparse and very tangled hierarchy by Noy and Hafner [11]. Manual building of large ontologies, such as Cyc Knowledge Base, demands a substantial amount of human efforts, which is the reason that all domains are not covered yet in all details. Further extension of such a large ontology is

* Corresponding author. Tel.: +386 41298345; fax: +386 14773315.
E-mail address: inna.koval@ijs.si (I. Novalija).

0950-7051/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.knosys.2011.06.002

challenging as well because of its complexity and interconnectivity. Presented methodology is meant to fasten up the process of building an extensive ontology and lower the price of doing it.

The main contribution of this paper is in proposing a methodology for text-driven semi-automatic ontology extension using ontology content, ontology structure information and co-occurrence data between existing and candidate ontology concepts.

Ontology content of a particular concept is defined as the available textual representation of the referred concept. The ontology content includes a natural language concept denotation (lexical entries for a particular concept) and textual comments about the concept. **Ontology structure** of a particular concept is defined as the neighborhood concepts involved in the hierarchical and non-hierarchical relations with the referred concept. **Co-occurrence** information is represented by the occurrence of two or more concepts within a defined textual block. The available textual information is used to find the co-occurrences between existing ontology concepts and new domain concepts suggested for ontology extension.

Ontology extension in this paper stands for: adding new concepts to the existing ontology or, augmentation of the existing textual representation of the relevant concepts present in the ontology with new available textual information – extension of the concept comments, changing or adding concept denotation.

In experiments the suggested methodology for text-driven ontology extension, aggregating the elements of text mining and user interaction approach for ontology extension, is used for inserting the new knowledge into Cyc [12], which maintains one of the most extensive common-sense knowledge bases worldwide.

The experiments are performed in two different domains having two knowledge representation levels – financial domain represented by the glossary of financial terms [13] and fisheries & aquaculture domain represented by Aquatic Sciences and Fisheries Abstracts (ASFA) thesaurus [14].

The evaluation of the methodology for ontology extension shows its ability to expedite the ontology extension process. The experimental results demonstrate that using a combination of ontology content, structure and co-occurrence information is more beneficial for the extension of large multi-domain ontologies, than using only content, only co-occurrence or only concept denotation information.

Comparison of **OntoPlus** methodology with a number of ontology extension methods described below shows, that the presented methodology overcomes the weaknesses of other approaches. **OntoPlus** methodology is based on the ontology contextual information; it is suitable to very large multi-domain ontologies; it utilizes the language independent approaches; it allows for transforming textual information organized at different knowledge representation levels into a structured conceptualized form. From the hundreds thousands of concepts, the proposed methodology is able to find the concepts and relationships the user needs and present them in a ranked list based on their relevance. The utilization of the lexical and structural information of the extensive knowledge bases and ontologies contributes to their infinite extension and reuse.

The paper is structured as follows: Section 2 presents the overview of existing approaches to ontology extension; problem definition is given in Section 3; the new methodology for ontology extension is discussed in Section 4; this section also contains the adaptation of the proposed methodology for one concrete scenario – Cyc Knowledge Base extension; Section 5 describes the evaluation – experiments and results; the discussion is covered in Section 6, and finally, we present a conclusion in Section 7.

2. Related work

The automatic and semi-automatic ontology extension processes are usually composed of several phases. Most approaches include defining the set of the relevant ontology extension sources, preprocessing the input material, ontology augmentation according to the chosen methodology, ontology evaluation and revision phases.

Buitelaar et al. [15] state that the process of ontology development from text can be organized in a layer cake of increasingly complex subtasks: terms extraction at the bottom, synonyms extraction, concepts definition, establishment of concept hierarchies, relations identification and rules definition on the top. As Reinberger and Spyns [16] state, the following steps can be found in the majority of methods for ontology learning from text: collecting, selecting and preprocessing of an appropriate corpus, discovering sets of equivalent words and expressions, establishing concepts with the help of the domain experts, discovering sets of semantic relations and extending the sets of equivalent words and expressions, validating the relations and extended concept definition with help of the domain experts and creating a formal representation. As suggested in [17], ontology learning from text is just one phase in the methodology for semi-automatic ontology construction preceded by domain understanding, data understanding and task definition and followed by ontology evaluation and ontology refinement. In this paper, we focus on ontology extension assuming that the main challenge is in finding the relevant concepts and relations in the existing ontology.

Natural language processing is notably used for learning or extending ontologies [18,19]. Unsupervised text mining for ontology learning was elaborated by Reinberger and Spyns [16]. Cimiano et al. [20] suggest an approach for learning concept hierarchies from text based on Formal Concept Analysis (FCA), a method mainly used for the data analysis. The Web is considered a source of text suitable for ontology extension in [21], where the English lexical ontology WordNet [22] is extended based on clustering word senses. However, our approach is more general by enabling extension of any ontology that has some lexical description of the concepts.

Prieto-Diaz [23] utilizes top-down and bottom-up processes for ontology development. A more general top-down process embodies domain experts identifying the key concepts in order to capture the high level ontology. The instruments for the text analysis are used in the bottom-up process for keywords extraction. In a similar way, our methodology incorporates top-down and bottom-up process, where the user is providing relevant keywords or glossary while the system uses the data to identify relevant parts of the existing ontology.

Lexico-syntactic pattern-based ontology learning is handled by Text2Onto [24], a framework for ontology learning and data-driven change discovery. The main aspects of the Text2Onto framework include using so called Probabilistic Ontology Model, user interaction and operation strategies for data-driven change discovery. Text2Onto allows learning ontological structures from text in form of modeling primitives, such as concepts, subclasses, instances etc. without connection to a certain representation language. SPRAT [25] is a tool for automatic semantic pattern-based ontology population. SPRAT system combines the name entity recognition, ontology-based information extraction and relation extraction in order to define patterns for the identification of a variety of entity types and relations between them. In our work, patterns are not utilized, as we are assuming availability of keywords or glossary terms that already represent new concepts used for ontology extension.

Fortuna et al. [26] developed an approach to semi-automatic data-driven ontology construction focused on topic ontology. The approach combines machine learning and text mining techniques with an efficient user interface. The domain of interest is described by keywords or a document collection and used to guide the ontol-

ogy construction. OntoGen [26] uses the vector-space model for document representation. The tool operates based on a cosine similarity between textual documents.

Several methods of automatic ontology extension operate with enlarging of Cyc Knowledge Base. The automated population of Cyc with named entities involves the Web and a framework for validating candidate facts [27]. The semi-automatic approach for Cyc KB extension presented in [28] is based on the user-interactive dialogue system for knowledge acquisition, where, the user is engaged in a natural-language mixed-initiative dialogue. The system contains a natural language generation module, parsing module, post-processing module, dictionary assistant, user interaction agenda and salient descriptor. Medelyan and Legg [29] describe the methodology for integrating Cyc and Wikipedia, where the concepts from Cyc are mapped onto Wikipedia articles describing correspondent concepts. Sarjant et al. [30] use Medelyan and Legg [29] method to augment Cyc ontology using pattern matching and link analysis. In our approach Cyc is used as an example of a large ontology on which our methodology can be applied, and we currently assume that the list of keywords representing new concepts to be used in ontology extension is provided as such without a need for extracting them from free text. Simultaneously, comparing the **OntoPlus** methodology to the specific approaches, which deal only with Cyc KB extension, we find that our approach is constructed to be more general and can be applied to any other ontology, or any other domain.

Extension of the existing ontology by automatically extending its relations was addressed by several researchers. The approaches include learning taxonomic [31]/non-taxonomic relations [32] and extracting semantic relations from text based on collocations [33]. However, in the presented work we do not address the problem of extending set of relations, but assume suggesting relevant existing relation instances.

Turney [34] has used a co-occurrence analysis technique for mining synonyms from Web. Besides, ontology structure has been adequately used in the collective entity resolution [35]. Usage of co-occurrence in our methodology was inspired by the work on collective entity resolution and synonym extraction.

3. Problem definition

In this Section we describe the formal background of the proposed methodology. According to Maedche and Staab [36], **Ontology (*O*)** is a tuple:

$$O := \{L, C, H_c, R, H_r, F, G, A\} \tag{1}$$

where L represents lexical entries for concepts and relations; C is a set of concepts; H_c is a taxonomy of concepts; R is a set of non-taxonomic relations; H_r is a set of taxonomic relations; F and G are the relations connecting concepts and relations with lexical entries from L; A is a set of axioms.

Following the ontology definition, it is possible to formalize the ontology extension problem in a similar way to the ontology learning task [26]:

$$f : (O, T) \rightarrow O^e \tag{2}$$

O represents an existing ontology, which we are extending; T is a domain glossary – textual source of information we use for ontology extension; O^e is an extended ontology.

The proposed methodology enables extending the existing ontology by (a) adding a new hierarchically related concept, by (b) augmenting textual representation of the existing concept or, by (c) adding new axioms.

(a) The following formula corresponds to adding a new hierarchically related concept to the existing ontology $O(1)$:

$$f_{HRC} : (L, C, H_c, R, H_r, F, G, A) \rightarrow (L \cup \{l\}, C \cup \{c\}, H_c^e, R, H_r, F^e, G, A) \tag{3}$$

c is a new hierarchically related concept; l is a lexical entry for a new hierarchically related concept; H_c^e s an extended taxonomy of concepts; $Fe$is an extended set of relations connecting concepts with lexical entries from L.

(b) Augmentation of the existing textual representation of the relevant concepts of the existing ontology $O(1)$ with new lexical entries is displayed as:

$$f_{LE} : (L, C, H_c, R, H_r, F, G, A) \rightarrow (L \cup \{l\}, C, H_c, R, H_r, F^e, G, A) \tag{4}$$

l is a lexical entry for an existing concept; F^e is an extended set of relations connecting concepts with lexical entries from L.

(c) Adding a new axiom to the existing ontology $O(1)$ is presented in the following way:

$$f_A : (L, C, H_c, R, H_r, F, G, A) \rightarrow (L, C, H_c, R, H_r, F, G, A^e) \tag{5}$$

A^e is an extended axiom set.

4. Methodology

We propose a new **OntoPlus** methodology for text-driven ontology extension, which combines text mining methods with user-oriented approach and supports the extension of multi-domain ontologies.

The detailed description of seven methodology phases and application of the methodology for extension of Cyc Knowledge Base are given below.

The proposed methodology embodies three main modules: the Domain Information Module (DIM), the Domain Subset Extraction Module (DSEM) and the Ontology Extension Module (OEM). The methodology is mainly targeted at the ontology engineers – people, who develop and maintain large ontologies, such as Cyc KB.

The main task of the Domain Information Module is accumulating the relevant domain information, needed for the ontology extension. The Domain Information Module contains the domain keywords, determined by the user and a domain relevant glossary of terms with descriptions.

In the Domain Subset Extraction Module initially the multi-domain ontology is limited to the particular domains of interest. Subsequently, the domain related knowledge is extracted from the ontology.

The Ontology Extension Module is the most important methodology module where the actual procedure of ontology extension takes place. For each glossary term OEM produces the ranked list of textually related existing ontology concepts. In addition, the relationship for every term-concept pair is suggested. Based on the suggested list of related concepts with relationships, the user makes a decision of which terms from the domain relevant glossary should be added to the ontology and how the glossary terms should be connected with the existing ontology concepts. Ontology is extended after the user validation. The technical aspects of the proposed methodology are described in the remaining of this section.

In detail, the proposed methodology for text-driven ontology extension accounts for the following phases:

1. *Domain information identification.* The domain information identification is taking place in the Domain Information Module.

The user identifies the appropriate domain keywords. As well, in this module a domain relevant glossary, containing terms with descriptions is determined. We assume that the glossary terms are the candidate entry concepts for the existing ontology. Consequently, the glossary terms might be in the following relationships with the existing ontology concepts:

- Equivalence relationship: candidate concept represented by a glossary term is equivalent to the existing ontology concept.
- Hierarchical relationship: candidate concept represented by a glossary term is in the superclass-subclass relationship with an existing ontology concept.
- Non-hierarchical relationship: candidate concept represented by a glossary term is the in the associative relationship with an existing ontology term. The nature of the relationship is not hierarchical.
- No relationship: candidate concept represented by a glossary term is not related to the existing ontology concept.

2. *Extraction of the relevant domain ontology subset from multi-domain ontology.* Extraction of the relevant domain ontology subset from multi-domain ontology based on the specified domain information is taking place in the Domain Subset Extraction Module. In case of large common-sense ontologies, such as Cyc Knowledge Base, the user entering new knowledge very often needs a particular ontology subset of his domain interest. Therefore, the domain keywords are mapped to the natural language representation of the ontology domain information and a set of the relevant domains of interest is identified.

Further, ontology concepts defined in these domains are extracted. By concept extracting we mean obtaining the content and structure of the ontology concept. Correspondently, we find the textual representation (natural language denotation and comments) as content for the particular ontology concept. The ontology structure of the particular concept is represented by the natural language denotations of the hierarchically and non-hierarchically connected ontology concepts. Besides that, the names of the glossary terms are mapped to the natural language denotations of the concepts from other domains and the correspondent concepts are also extracted.

Fig. 1 demonstrates the extraction of Business & Finances knowledge subset from the multi-domain ontology. The ontology contains three domains – Business & Finances, Transportation & Logistics and Politics.

The domain relevant glossary is composed of the financial terms with descriptions. The correspondent financial concepts are extracted from the Business & Finances domain defined by the user-specified keywords. Moreover, concepts from other domains with concept denotations equivalent to the glossary term names are extracted.

3. *Domain relevant information preprocessing.* The information from the domain relevant glossary and the extracted relevant ontology subset are linguistically preprocessed in the Ontology Extension Module. The preprocessing phase includes tokenization, stop-word removal and stemming. A chain of linguistic components, such as tokenization, stop-word removal and stemming allows normalizing the textual representation of ontology concepts and a domain relevant glossary of terms with their descriptions. Textual information is represented using bag-of-words representation with normalized TFIDF weighting and similarity between two text segments is calculated using cosine similarity between their bag-of-words representations, as commonly used in text mining [17]. For each term from the domain relevant glossary we compose bag-of-words aggregating preprocessed textual information from: (a) the glossary term name and (b) the term comment. For each concept from the extracted relevant ontology subset the following information is considered: (a) the ontology concept content consisting of the preprocessed natural language concept denotation and concept comment; (b) the ontology concept structure consisting of the preprocessed natural language concept denotation and natural language denotations of hierarchically and non-hierarchically related concepts.

Ontology concept denotation and ontology concept comment carry the same semantic weight. As well, we attribute the same semantic weight to the glossary term name and glossary term comment.

In addition, for relation identification, for each ontology concept we compose two additional bags-of-words: one with natural language denotation of the concept and natural language denotations of superclasses of this concept, another with natural language denotation of the concept and natural language denotations of subclasses of this concept.

4. *Composing the list of potential concepts and relationships for ontology extension.* The ranked list of the relevant concepts and possible relationships suitable for ontology extension is composed in this phase. Cosine similarity $similarity_{content}(t, c)$ between glossary term t and ontology concept c content is calculated and weighted with weight δ_1 defined by the user. Cosine similarity $similarity_{structure}(t, c)$ between glossary term t and ontology concept c structure is calculated and weighted with weight δ_2.

The complexity of calculating content similarity and structure similarity between pairs of ontology concepts and glossary terms is $O(mn)$, where m is the number of glossary terms and n is the number of ontology concepts.

We use Jaccard similarity to measure the co-occurrence of glossary term t and ontology concept c:

$$similarity_{co\text{-}occur}(t, c) := \frac{N(t, c)}{N(t) + N(c) - N(t, c)} \tag{6}$$

$N(t, c)$ is the number of textual documents where glossary term t and ontology concept c occur together. $N(t)$ is the number of documents where glossary term t occurs and $N(c)$ corresponds to the number of documents which contain ontology concept c. Co-occurrence similarity is calculated based on the names of glossary terms and ontology concepts denotations. Each textual document is composed either of the content of an ontology concept or of the textual information about a particular glossary term (name and description). The combined content, structure and co-occurrence similarity $similarity(t, c)$ is used to rank ontology concepts for each glossary term:

$$\begin{aligned} similarity(t, c) &:= \delta_1 * similarity_{content}(t, c) + \delta_2 \\ &\quad * similarity_{structure}(t, c) + \delta_3 * similarity_{co\text{-}occur}(t, c) \\ \sum_i \delta_i &= 1, \quad i \in [1 \ldots 3] \end{aligned} \tag{7}$$

Ontology concepts with similarity $similarity(t, c)$ larger than $similarity_{max}(t, c^{max})*(1 - \beta)$ are suggested to the user, where $similarity_{max}(t, c^{max})$ represents the highest similarity value between ontology concepts and a glossary term t and β is a user defined parameter:

$$similarity(t, c) \geqslant similarity_{max}(t, c^{max}) * (1 - \beta) \quad 0 \leqslant \beta \leqslant 1 \tag{8}$$

To propose the relationship of equivalence we use string-edit distance between glossary term names and the related concept names. In the case of equivalence, the user can extend textual representation of the related ontology concept. According to formula (4), the user is able to insert into the ontology $O(1)$ the new lexical entry l, obtained from the glossary term t lexical information:

$$O^e := \{L \cup \{l\}, C, H_c, R, H_r, F^e, G, A\} \tag{9}$$

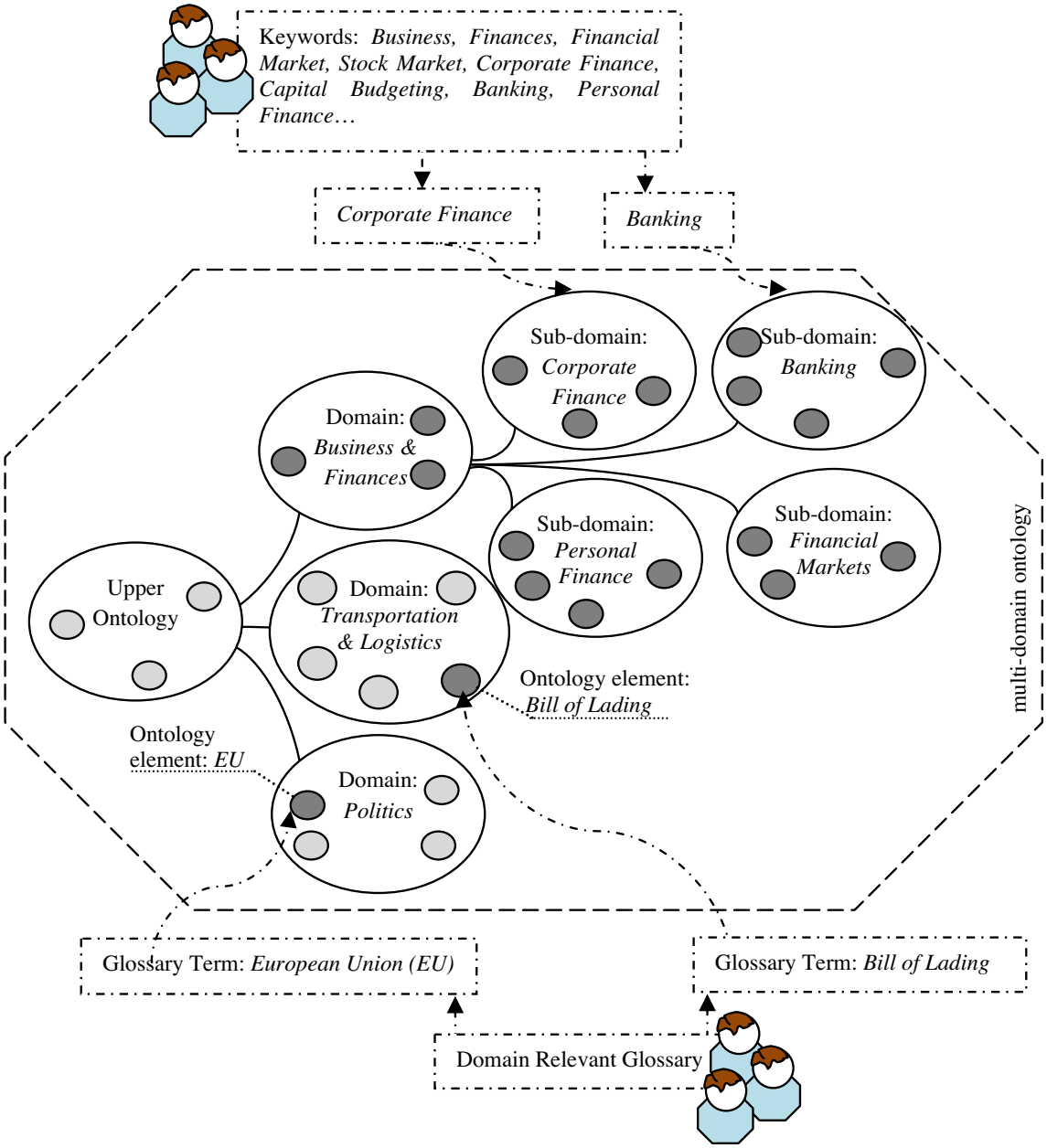

Fig. 1. Illustrative extraction of Business and Finances domain subset from a multi-domain ontology (dark circles represent the extracted relevant concepts).

The observations show that the name and comments of the glossary terms often contain references to the superclasses and subclasses of the related ontological concepts.

Hierarchical relationship: *"glossary term t is a subclass for concept c"* is proposed, when the similarity $similarity_{sub}(t, c)$ between the glossary term t and subclasses of the related concept c is higher than the similarity $similarity_{sup}(t, c)$ between the glossary term t and superclasses of the related concept c; γ is a user defined parameter:

$$\frac{similarity_{sub}(t,c)}{similarity_{sup}(t,c)} \geqslant \gamma + 1 \quad 0 \leqslant \gamma \leqslant 1 \tag{10}$$

We rank the hierarchical (subclass) relations using the quotient in the left part of formula (10).

Adding a new subclass relationship to the ontology $O(1)$ using formula (3), we get t as a new concept hierarchically related to the existing ontology concept c; l as a lexical entry for a new hierarchically related concept; H_c^e as an extended taxonomy of concepts and F^e as an extended set of relations connecting concepts with lexical entries from L:

$$O^e := \{L \cup \{l\}, C \cup \{t\}, H_c^e, R, H_r, F^e, G, A\} \tag{11}$$

Currently we do not propose hierarchical relationship *"glossary term t is a superclass of concept c"* since we assume that the existing ontology concepts, which are already embedded into the hierarchy, contain a valid superclass information. If we do not find equivalence or hierarchical relationships between glossary term t and the related concept c or if the nature of the relationship is not clear (for instance, when the related ontology concept has no subclasses), we propose non-hierarchical associative relationship: *"glossary term t is conceptually related to ontology concept c"*.

Using formula (5) we obtain in the ontology an extended axiom set A^e:

$$O^e := \{L, C, H_c, R, H_r, F, G, A^e\} \quad (12)$$

5. *User validation.* Furthermore, in OEM the user validates the candidate entries results consisting of the glossary terms, existing ontology concepts and glossary term-ontology concept relationships. In case of the equivalence relationship the user can extend the textual representation of the existing ontology concept by adding comment, adding or changing the natural language denotation. In case of the hierarchical relationships the user can add subclasses to the existing ontology concepts. If the nature of the relationship is not clear, the user can create an associative relationship or choose any other relationship between a glossary term and existing ontology concept. Moreover, the list with validated entries in the relevant format is created. The experiments show that limiting the number of results and presenting them to the user for validation allow achieving a high level quality control.

6. *Ontology extension.* The ontology extension is taking place in the Ontology Extension Module. It represents adding the new concepts and relationships between concepts into the ontology.

7. *Ontology reuse.* The ontology reuse phase serves as the connection link between separate ontology extension processes. As a part of the new extension process, we reuse the previously extended ontology in the Domain Subset Extraction Module and in the Ontology Extension Module.

Each phase of the methodology workflow described above is intended to fasten the process of ontology extension. *Domain information identification* and *Extraction of the relevant domain ontology subset from multi-domain ontology* help to restrict the area of ontology extension to a specific domain, so the users deal only with their sphere of interest. *Domain relevant information preprocessing* is a necessary act for identification of the related ontology concepts and correspondent relationships. The experiments performed on Cyc Knowledge Base applying **OntoPlus** methodology justify that the combination of ontology content, ontology structure and co-occurrence information provides the user with high number of concepts suitable for building an ontology, what expedites the process of the ontology extension. The experiments demonstrate that the content, structure and co-occurrence weight may vary between domains and knowledge representation levels.

4.1. Extension of Cyc Knowledge Base

We have adapted the proposed methodology in order to obtain an exhaustive specific methodology for Cyc Knowledge Base extension.

Currently, Cyc operates on one of the largest knowledge bases in the contemporary IT world. New assertions are continually added to Cyc KB manually. In addition, term-denoting functions allow for the automatic creation of millions of non-atomic terms and Cyc adds an enormous number of assertions to the KB automatically as a product of the inference process [9]. The contents of the Cyc KB are represented in CycL [37], a formal language based on second order logic.

Fig. 2 displays the Cyc adaptation of the proposed methodology for semi-automatic ontology extension.

The methodology phases are illustrated with numbers. The main adaptations compared to the methodology described above are based on microtheories (Mt) [38] that Cyc is using to represent thematic subsets of the ontology.

Namely, the knowledge base in Cyc is divided into various microtheories which contain a set of facts valid in a particular context. The graphical representation of the ontology extension process, shown in Fig. 2, demonstrates semi-automatic adding of new concepts to Cyc Knowledge Base.

In the *Domain information identification phase (1)* in the Cyc adaptation of the methodology for ontology extension we identify the Domain Keywords and the Domain Glossary for the domains of interest. For the research purposes we use a financial glossary, composed by Harvey [13] and ASFA thesaurus [14].

The Relevant Ontology (Cyc KB) Subset is extracted in the *Domain Subset Extraction Module*. The Upper-Level Domain Extractor uses Domain Keywords to obtain a number of domain relevant Cyc microtheories by mapping microtheory names with domain keywords from the list. Furthermore, the Knowledge Extractor provides a set of concepts defined in the domain relevant microtheories. Additionally, the concepts that are defined in other microtheories, but contain the natural language denotations correspondent to glossary term names, are extracted into the Relevant Ontology (Cyc KB) Subset.

As mentioned above, the role of the Domain information identification phase (1) and Extraction of the relevant domain ontology subset from multi-domain ontology phase (2) in the proposed workflow consists in limiting the ontology extension process to a specific domain of interest. The very large size of Cyc KB does not allow for an effective and fast related concept search without such domain restriction. In addition, it allows the users to deal only with their sphere of interest.

The Domain relevant information preprocessing phase (3) and the Composing the list of potential concepts and relationships for ontology extension (4) follow subsequently. The Ontology Extender takes the Domain Glossary and extracted Cyc concepts as an input. The bag-of-words containing term name and term description is composed for each glossary term. A set of bag-of-words is composed for every extracted Cyc concept: concept denotation and concept comment; concept denotation, denotations of concepts in the hierarchical and non-hierarchical relationships with extracted Cyc concept; concept denotation and denotations of superclasses of the extracted concept; concept denotation and denotations of subclasses of the extracted concept.

The *Domain relevant information preprocessing (3)* is included into the workflow as a standard ontology learning step [24–26]. In order to find the related Cyc concepts for each glossary term we use TF-IDF weight and cosine similarity for content and structure similarities and Jaccard similarity for co-occurrence similarity.

The utilization of the text mining methods in the proposed methodology and combining ontology content, structure and co-occurrence information allows us to provide the user with higher number of concepts suitable for the ontology extension than using only concept denotations, only ontology content or only co-occurrence analysis.

Our experiments show that the best results are obtained giving more weight to content and structure for the financial domain and more weight to content and co-occurrence for fisheries & aquaculture domain. Cyc concepts with combined similarity larger than $similarity_{max}(t, c^{max}) * (1 - \beta)$, where $similarity_{max}(t, c^{max})$ represents the maximum combined similarity value between Cyc concepts and a glossary term t for a particular glossary term t, are suggested to the user. Including β parameter into the model allows us to restrict the number of concepts presented to the user. The experiments confirm that in both domains higher values of β lead to the higher hit rates, but at the same time, the user should check larger amount of the suggested related Cyc concepts.

We use string-edit distance between glossary term names and related concept denotations to propose the relationship of equivalence. In this case the user can extend Cyc textual representation of the related concept – add information to Cyc comment, add or change the Cyc concept denotation.

In case the similarity between the glossary term t and subclasses of the related Cyc concept c is higher than the similarity between the glossary term t and superclasses of the related Cyc concept c by γ, we propose hierarchical relationship: "*glossary term t is a subclass for Cyc concept c*". Including γ parameter into the

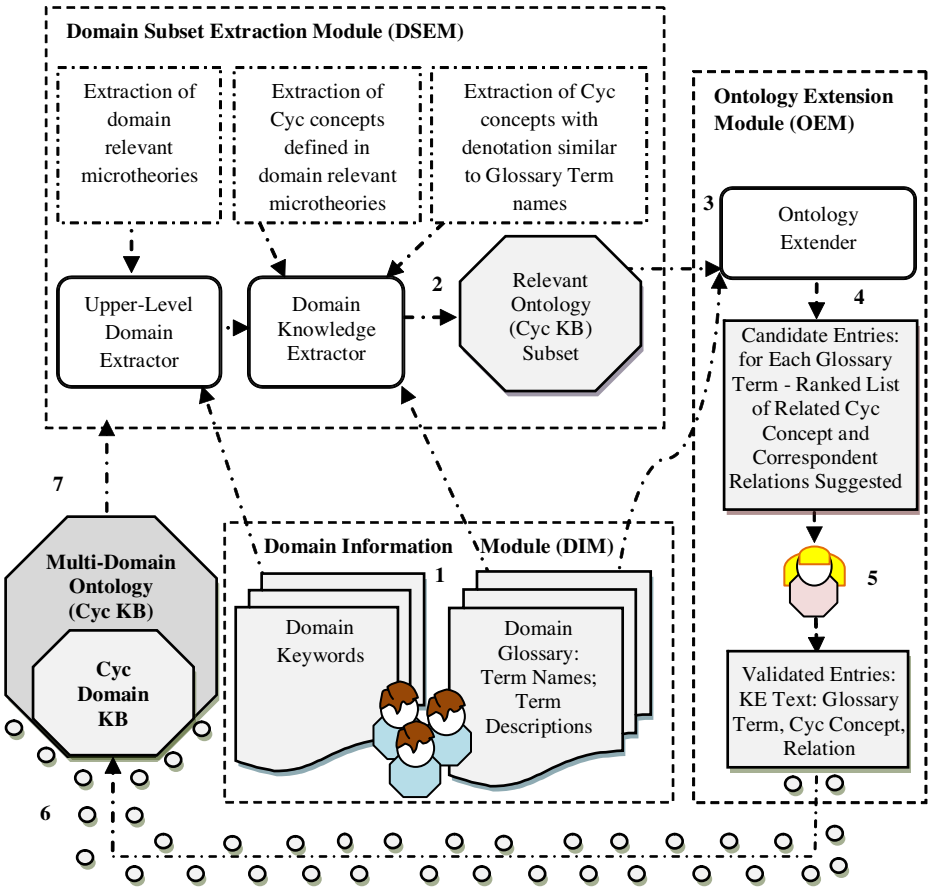

Fig. 2. Text-driven ontology extension (Cyc KB adaptation)

model allows us to define the number of relationships presented to the user. The experiments confirm that in both domains higher value of γ leads to smaller number of proposed relationships. At the same time, both γ and β parameters influence the learning accuracy measure.

If we do not find equivalence or hierarchical relationships between glossary term t and related Cyc concept c, we propose non-hierarchical associative relationship: *"glossary term t is conceptually related to Cyc concept c"*.

Evaluation of the relation identification shows that using the proposed methodology, we can provide the user with correct automatically suggested equivalent, hierarchical (subclass) and associative relationships.

In the *User validation phase (5)* the user has a possibility to dynamically construct new assertions in the Knowledge Entry (KE) format which can be then automatically integrated into Cyc KB.

Afterwards, Cyc KB extension is taking place in the *Ontology extension phase (6)*. The *Ontology extension phase (6)* is the actual point of the workflow, where the Cyc KB is extended with new concepts, textual information about existing concepts and relationships between concepts. The *Ontology reuse phase (7)* occurs when the new set of knowledge is added to Cyc KB. Since ontology extension is a repetitive process, we include this phase as a final stage of the methodology.

5. Evaluation

In order to evaluate the proposed methodology we have conducted a series of experiments on the data sources, described below. The experiments are conducted in two domains: financial domain and fisheries & aquaculture domain.

For the proposed methodology evaluation we have used two evaluation techniques – the manual evaluation by human experts and gold standard based approach [39].

The evaluation is performed at the lexical, taxonomic (concept hierarchy) and non-taxonomic levels. For the lexical evaluation the mapping of the glossary terms to the existent ontology concepts is performed. At the taxonomic level the evaluation of the suggested hierarchically related concepts and suggested superclass-subclass relationships is implemented. Finally, at the non taxonomic relations level the evaluation of the suggested associatively related concepts and associative relationships is done. While the gold standard based approach is used to perform lexical and taxonomic evaluation, the manual evaluation is used at the non-taxonomic level.

Maedche and Staab [40] have used the normalized string edit distance to identifying how similar two ontologies are. Normalized string edit distance between ontology concept denotations and glossary term names is used as a baseline measure in the evaluation of the proposed **OntoPlus** methodology.

In order to define how successful the proposed methodology is in practice, we have used the evaluation measures commonly used for ontology learning evaluation.

Precision of the top suggested concept defines the percentage of the glossary terms for which the equivalent and hierarchical, associative or any related ontology concepts have obtained the highest position in the suggested ranked related concept list:

$$Precision := \frac{TP}{TP + FP} \tag{13}$$

where *TP* represents the correct related concepts identified; *FP* represents the false related concepts identified.

Learning accuracy [41] shows the degree to which the proposed methodology correctly predicts the superclass for the candidate ontology concept (represented by a glossary term) to be learned:

$$LA := \sum_{i\in\{1\ldots n\}} \frac{LA_i}{n} \tag{14}$$

$$LA_i := \begin{cases} \frac{CP_i}{SP_i} & \text{if } FP_i = 0 \\ \frac{CP_i}{FP_i+DP_i} & \text{if } FP_i \neq 0 \end{cases} \tag{15}$$

where n represents the number of concept hypotheses for the target; SP_i is the length of the shortest path from the top node of the concept hierarchy to the maximally specific concept subsuming the instance to be learned in hypothesis i; CP_i is the length of the path from the top node to that concept node in hypothesis i which is common both to the shortest path (as defined above) and the actual path to the predicted concept (whether correct or not); FP_i is the length of the path from the top node to the predicted false; DP_i is the node distance between the predicted false node and the most specific common concept still correctly subsuming the target in hypothesis i.

In addition, we have used a **hit rate** measure used in the evaluation of recommendation systems. The hit rate displays the number of hits and their position within top N suggestions [42]. We specify the hit rate measure as following:

$$HR := \frac{\sum_{t\in T} HR_t}{|T|} \tag{16}$$

$$HR_t := \frac{\sum_{u\in U} HIT(t,u)}{|U|} \tag{17}$$

where t is a candidate concept for ontology extension; u represents a user; $HIT(t,u)$ is a binary function. For the candidate concept t and user u it returns 1 if the correspondent related ontology concepts have been found among the top N suggestions and 0 otherwise; U is a set of users; T represents the set of candidate ontology concepts (glossary terms).

5.1. Data description

According to the first phase of our methodology, domain knowledge identification should be made in the initial phase. For the financial domain, we have selected the Harvey [13] financial glossary which can be found at the Yahoo! Finance website [43]. The Harvey financial glossary [13] contains around 6000 hyperlinked financial terms. The typical financial glossary entries are demonstrated in Fig. 3.

Fisheries & aquaculture domain, represented by the Aquatic Sciences and Fisheries Abstracts (ASFA) thesaurus [14], has been selected as a second domain of interest. ASFA thesaurus contains around 9900 terms involving several types of relationships: equivalence relationships (USE, Use For UF), hierarchical relationships (Broader Term BT, Narrower Term NT), associative relationships (Related Term RT) and notes (SN). Fig. 4 shows how BT, NT, RT, USE, UF and SN ASFA thesaurus relationships are used to compose a COMMENT for a particular term in the fisheries & aquatic thesaurus, in order to get the equivalent content as provided by a glossary (term and its comment/description).

5.2. Experimental settings

In order to evaluate the suggested methodology, we have conducted a number of experiments on a subset of 100 randomly selected terms from each domain resource (financial glossary, fisheries & aquaculture thesaurus). The line of experiments with examples in the financial domain is illustrated with numbers in Fig. 5.

Performing the domain information identification, apart from the domain relevant glossary, we have defined the domain keywords for the financial and fisheries & aquaculture domains. The human experts annotated the selected terms from the Harvey financial glossary [13] and ASFA thesaurus [14] with the correspondent equivalent and hierarchically related terms from Cyc Knowledge Base. The extensive size of Cyc Knowledge Base does not allow the experts to annotate the selected glossary term with all related concepts from the ontology.

We have performed the domain information preprocessing and extraction of the relevant domain ontology subset from Cyc Knowledge Base according to methodology phases described in Section 4. Using formulas (7), (8), and (10) from **OntoPlus** methodology, we are able to define a list of related Cyc concepts and a list of possible relationships for each glossary term. Cyc Knowledge Base is then extended with the concepts corresponding to the chosen terms based on the ranking proposed by the methodology.

We have used precision and hit rate measures to identify the importance of the ontology concept content, ontology concept structure and co-occurrence for establishing relatedness between glossary terms and ontology concepts. Subsequently, we have evaluated each measure by estimating the quality of concept ranking.

In addition, we have used learning accuracy to measure the quality of the hierarchical (subclass) relation identification.

5.3. Results

The results of the experiments confirm the applicability of the suggested methodology for ontology extension to Cyc Knowledge Base augmentation. We have organized results of the experiments into three groups: evaluation of the results of concept ranking, evaluation of the results of relation ranking and illustrative examples of Cyc KB extension are presented in the following subsections.

5.3.1. Concept ranking

According to our methodology, for each glossary term the user wants to add to the ontology, a ranked list of the related ontology concepts is created. As it was discussed in Section 4 (Methodology), we assume the following relationships between the glossary terms and existing ontology concepts:

TERM: Endowment
COMMENT: Gift of money or property to a specified institution for a specified purpose.

TERM: Recession
COMMENT: A temporary downturn in economic activity, usually indicated by two consecutive quarters of a falling GDP.

Fig. 3. Example financial glossary entries.

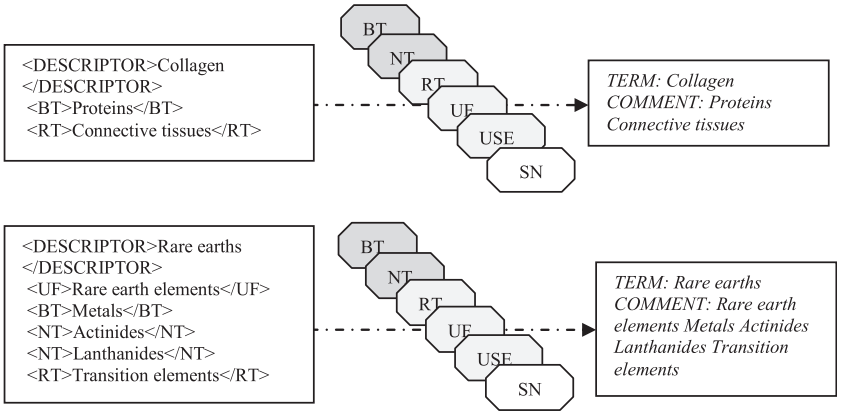

Fig. 4. ASFA thesaurus transformation.

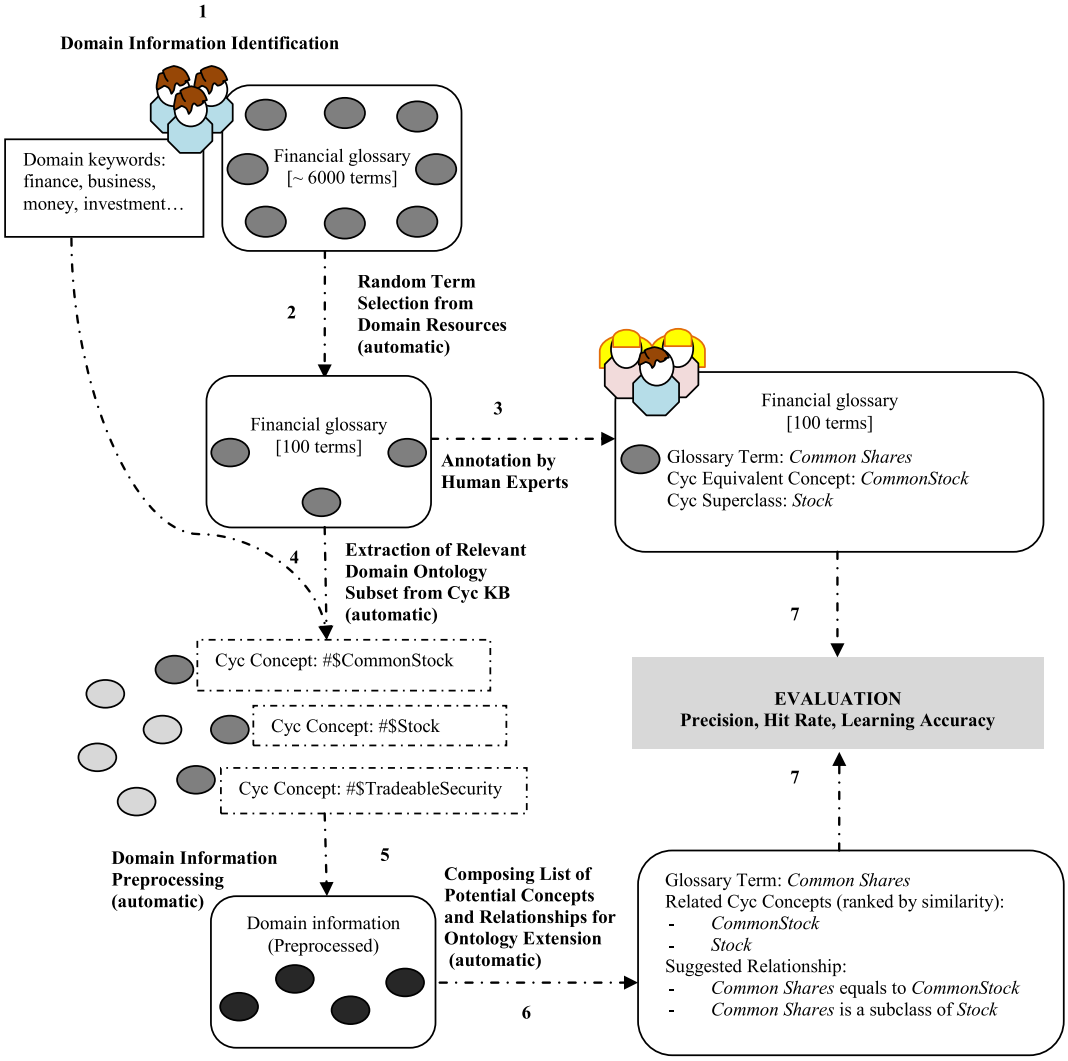

Fig. 5. Experimental settings diagram (financial domain).

- Equivalence relationship.
- Hierarchical relationship.
- Non-hierarchical (associative relationship).
- No relationship.

In the present experiment we have evaluated the quality of concept ranking depending on the different proportions of ontology concept textual content, ontology concept structure and co-occurrences of glossary terms and ontology concepts. Additionally, we have taken into the account the importance of the established relationship between the top suggested related ontology concept and candidate concept for the ontology extension. We have grouped equivalent and hierarchical relations in one group assuming that these relations are the most important in the ontology extension

process. Besides, we have also considered associative relations and a union of all the three considered relations (equivalent, hierarchical and associative relations) referred to as any relations.

Figs. 6 and 7 show the precision of the top one suggested concept depending on content, structure and co-occurrence information for the financial glossary and AFSA thesaurus, respectively. For instance, Fig. 6 provides the performance in the financial domain when only structure is used (the content weight and the co-occurrence weight are set to 0). Precision of the top ranked concept is 61% when analyzing any of the three considered relations (Any Rels), 30% when considering associative relations (Assoc Rels) and 31% when the top ranked concept is in the equivalent or hierarchical relations (Eqv and Hier Rels) with a correspondent glossary term.

For the financial glossary the best results are obtained by combining ontology content, ontology structure and co-occurrence information with giving more weight to content and structure and less weight co-occurrence. For ASFA thesaurus the co-occurrence plays a substantial role. In this case the best results are obtained by giving more weight to content and co-occurrence and less weight to structure. The potential explanation of such performance can refer to the different domain and structural nature of the financial glossary and fisheries & aquaculture thesaurus.

In addition, Tables 1 and 2 provide more detailed evaluation of the quality of ranking for the best performing weighting measures. The following weighting measures have been used for the financial domain: content weight $\delta_1 = 0.5$, structure weight $\delta_2 = 0.4$ and co-occurrence weight $\delta_3 = 0.1$. For fisheries & aquaculture domain we have set content weight $\delta_1 = 0.5$, structure weight $\delta_2 = 0.0$, co-occurrence weight $\delta_3 = 0.5$. Tables 1 and 2 contain the information on the hit rates (HR) for top 1, top 5 and top 10 suggested candidate concepts for ontology extension.

As a baseline measure we use mapping glossary term names to Cyc concept denotations, using normalized string-edit distance to rank the relations (equivalent or hierarchically related Cyc concepts and any related concepts) for each glossary term. Normalized string edit distance has been used for ontology learning and ontology matching purposes [40]. The string edit distance proposed by Levenshtein [44] measures the difference between strings by the smallest number of changes (insertions, deletions, substitutions) required for converting one string into another. We consider it a suitable baseline measure for the ontology extension problem since it produces the results even with the minimal information available, such as ontology concept denotations and glossary term names.

Furthermore, we have compared the best performing weighting measures with other methods, which use only cosine similarity between textual content of the documents [26] or only co-occurrence analysis [34].

The results of baseline are given under Baseline – Name [1.0] and show that by using baseline measure on both datasets for 40% of terms, the related Cyc concepts have been found among the top 10 suggested concepts (when considering any of the three relations – the last column in the tables marked as Any Rels). Comparing it to the results of the best performing combination of content, structure and co-occurrence (98% for financial domain and 96% for fisheries & aquatic domain, respectively), we demonstrate that by combining textual ontology content, ontology structure and co-occurrence information we can provide the user with more than double number of concepts suitable for the ontology extension than using the baseline.

In addition, we have performed a sensitivity analysis for β parameter (8). Fig. 8 displays the hit rate (HR) for the equivalent and hierarchically related concepts depending on β coefficient. It is possible to notice that higher hit rates both in financial and fishery & aquaculture domains are obtained with higher β values. Fig. 9 shows how β coefficient influences the cumulative number of concepts presented to the user. The lower β values imply the fewer number of the displayed concepts.

5.3.2. Relation ranking

In our methodology we have a possibility to suggest automatically not only the related existing ontology concepts, but also a relation: the equivalent, hierarchical (subclass) and associative

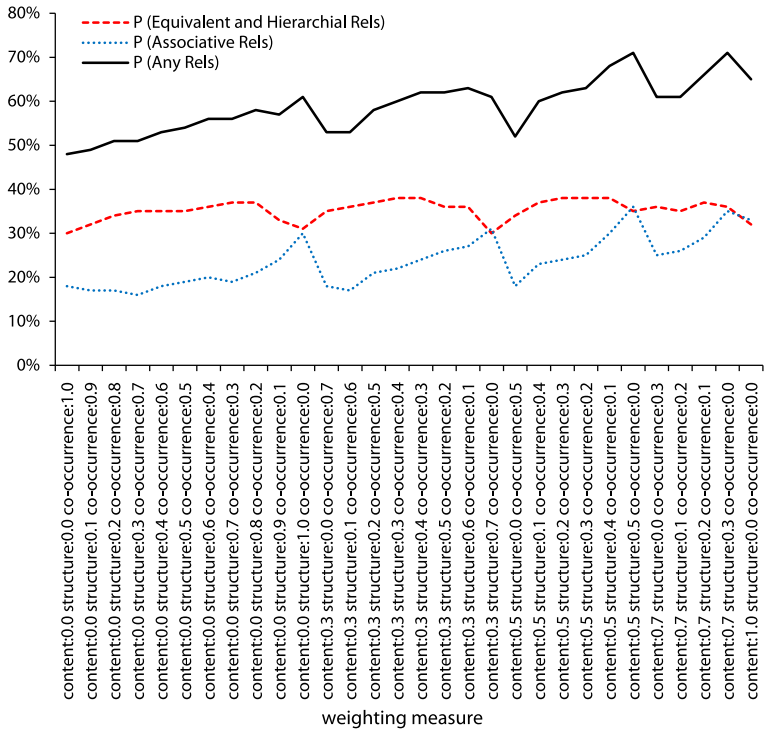

Fig. 6. Performance of the content, structure and co-occurrence weighting measures. Precision (P) of top 1 concept ranking (financial glossary).

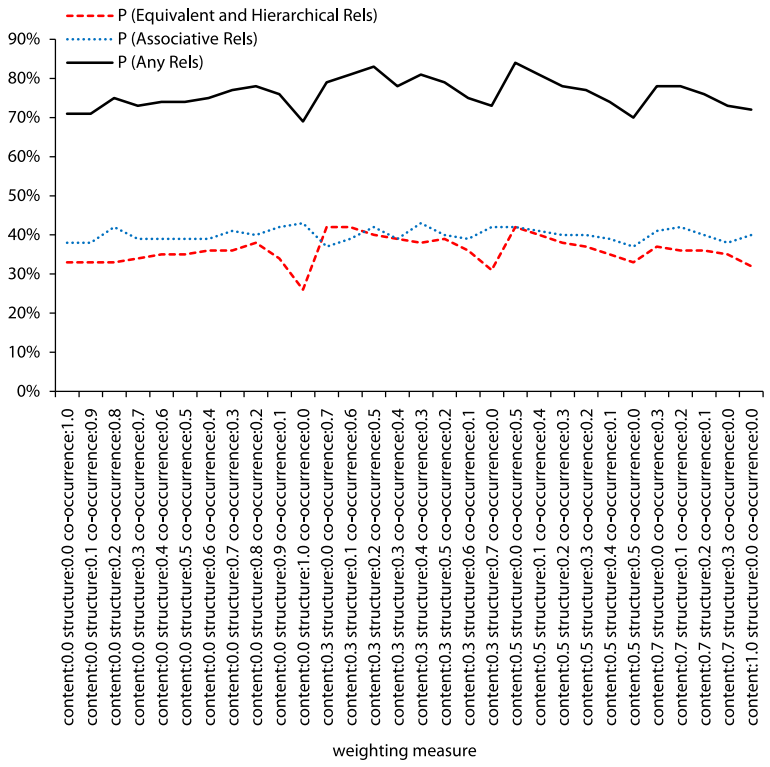

Fig. 7. Performance of the content, structure and co-occurrence weighting measures. Precision (*P*) of top 1 concept ranking (ASFA thesaurus).

Table 1
Evaluation of the top suggested candidate concepts for ontology extension (financial glossary).

Weighting measure		100 Random terms					
		HR (top 1)		HR (top 5)		HR (top 10)	
		Eqv. or hier. rels.	Any rels.	Eqv. or hier. rels.	Any rels.	Eqv. or hier. rels.	Any rels.
Baseline – Name	[1.0]	18	28	24	36	25	40
Content (cos. similarity)	[1.0]	32	65	60	92	68	95
Co-occur (Jaccard similarity)	[1.0]	30	48	48	62	52	73
Content Structure Co-occur	[0.5] [0.4] [0.1]	38	68	66	95	76	98

Table 2
Evaluation of the top suggested candidate concepts for ontology extension (ASFA thesaurus).

Weighting measure		100 Random terms					
		HR (top 1)		HR (top 5)		HR (top 10)	
		Eqv. or hier. rels.	Any rels.	Eqv. or hier. rels.	Any rels.	Eqv. or hier. rels.	Any rels.
Baseline – Name	[1.0]	24	37	25	38	27	40
Content (cos. similarity)	[1.0]	32	72	52	88	56	91
Co-occur (Jaccard similarity)	[1.0]	33	71	49	89	51	90
Content Structure Co-occur	[0.5] [0.0] [0.5]	42	84	63	96	66	96

relations between existing ontology concepts and candidate concepts for ontology extension.

For the relation identification experiment we have evaluated the precision of the suggested equivalent relations and a hierarchical (subclass) relation which obtained the highest position according to our methodology.

Table 3 displays the evaluation of the equivalent relations and top 1 subclass relations suggested for each candidate ontology concept.

The results in Table 3 show the precision (*P*) of 29.5% for subclass relations identification in financial domain and the precision of 20.3% for subclass relations identification in fisheries &

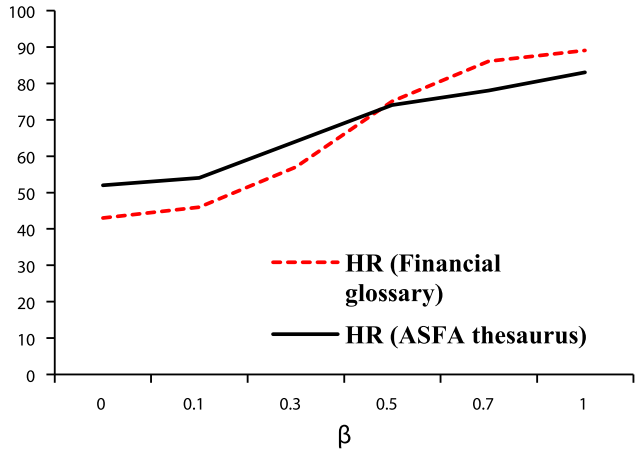

Fig. 8. Hit rate (HR) depending on β (equivalent and hierarchically related concepts).

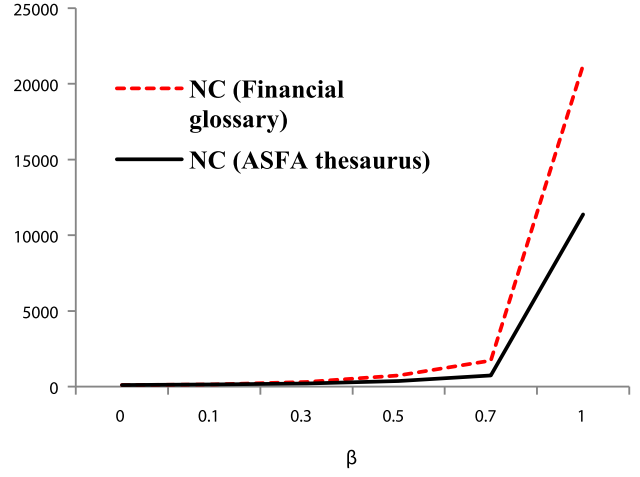

Fig. 9. Number of concepts (NC) depending on β.

aquaculture domain. The precision for equivalent relations identification is 60.0% for financial domain and 94.7% for fisheries & aquaculture domain.

In comparison, the authors of Text2Onto framework [24] report a precision of 17.38% for subclass-of relation identification on the subset of tourism-related texts. The evaluation in SPRAT [25] made on 25 randomly selected Wikipedia articles about animal shows the precision of 48.5% for subclass identification and 48.0% for synonym recognition.

The evaluation of the top suggested equivalent or hierarchical (subclass) relations is given in Table 4. The first column displays the number of concepts for which either equivalent or subclass relations, or both of them have been suggested automatically. Other three columns show the number of concepts for which the correct automatically suggested either equivalent or subclass relations, or both of them have been found among top 1, top 5 and top 10 suggested relations.

Evaluation of the suggested equivalent or hierarchical (subclass) relations shows that for 60 terms from the financial domain and for 41 concepts from the fisheries & aquaculture domain the correct automatically suggested relations have been found among top 10 suggested relations.

The sensitivity analysis for β and γ parameters is given on Figs. 10 and 11. Figs. 10 and 11 display the learning accuracy depending on $\beta(8)$ and $\gamma(10)$ for the candidate ontology concepts

Table 3
Evaluation of the equivalent, hierarchical (subclass) relations identification.

Glossary/weighting measure		100 Random terms	
		P (eqv. rels., %)	P (top 1 subclass rels., %)
Financial glossary		60	29.5
Content	[0.5]		
Structure	[0.4]		
Co-occur	[0.1]		
ASFA thesaurus		94.7	20.3
Content	[0.5]		
Structure	[0.0]		
Co-occur	[0.5]		

Table 4
Evaluation of the top suggested equivalent and hierarchical (subclass) relations.

Glossary/ weighting measure		100 Random terms			
		Number of concepts with eqv. or subclass rels. found	HR (top 1)	HR (top 5)	HR (top 10)
Financial glossary		97	36	56	60
Content	[0.5]				
Structure	[0.4]				
Co-occur	[0.1]				
ASFA thesaurus		80	31	40	41
Content	[0.5]				
Structure	[0.0]				
Co-occur	[0.5]				

from the financial glossary and ASFA thesaurus. It is possible to notice that the higher learning accuracy (LA) is obtained with $\gamma = 0.7$ ($0.0 \leqslant \beta \leqslant 1.0$) in the financial domain and $\gamma = 0.7$ ($0.0 \leqslant \beta \leqslant 0.3$) in fisheries & aquaculture domain.

Figs. 12 and 13 show the cumulative number of proposed relationships depending on β (8) and γ (10) for the candidate ontology concepts from the financial glossary and ASFA thesaurus. Both in the financial and fisheries & aquaculture domains higher γ values lead to the fewer number of relationships proposed to the user.

5.3.3. Examples of Cyc KB extension

Tables 5 and 6 provide the concrete examples of Cyc extension according to the proposed methodology.

The related Cyc concepts, which obtained the top position among the suggested related Cyc concepts and in the suggested equivalent, hierarchical (subclass) and associative relations, are highlighted in bold.

An example from Table 5 shows that using the proposed methodology and assuming that we would like to extend Cyc KB with a term *Life insurance* from the financial glossary, we get the composed list of the ranked related Cyc concepts:

- LifeInsurance.
- Endowment-LifeInsurance.
- InsurancePlan.
- FHAMortgageInsurance.
- VAMortgageInsurance.
- InsuranceClaimForm.
- MedicalInsuranceClaimForm.

After automatic relation identification the equivalent relation with Cyc concept *LifeInsurance*, hierarchical (subclass) relation with Cyc concept *InsurancePlan* and associative relations with Cyc concepts *Endowment-LifeInsurance*, *FHAMortgageInsurance*,

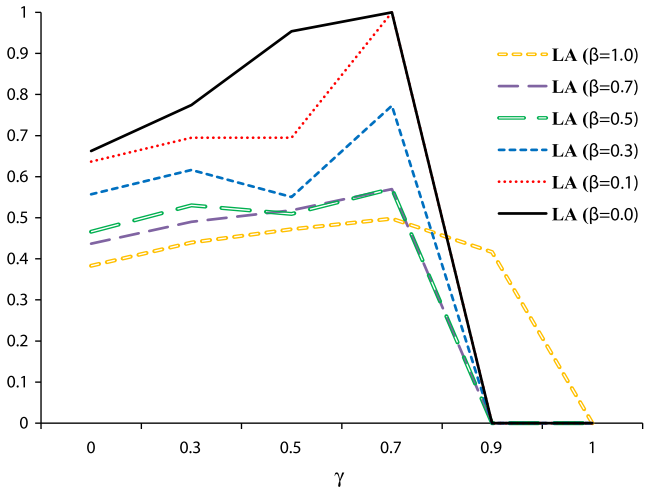

Fig. 10. Learning accuracy (LA) depending on β and γ (financial glossary).

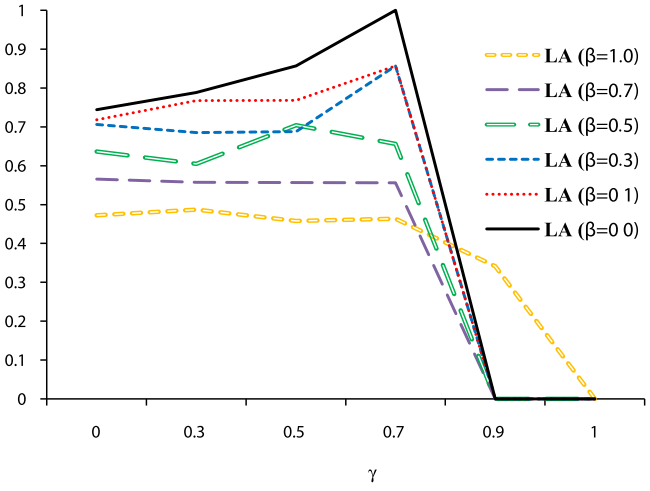

Fig. 11. Learning accuracy (LA) depending on β and γ (ASFA thesaurus).

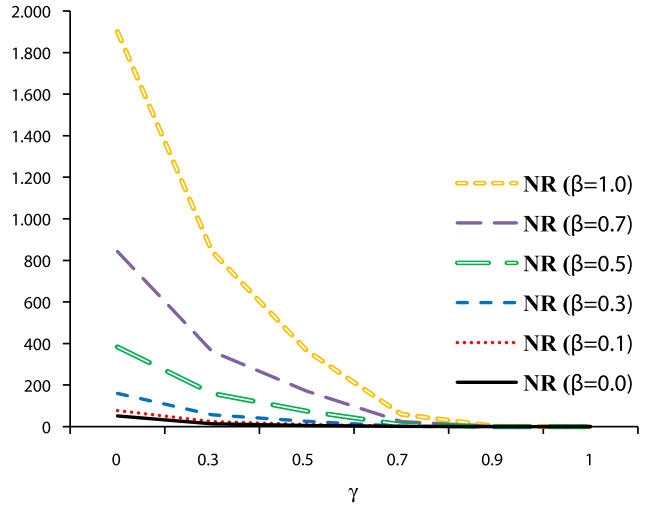

Fig. 12. Number of proposed relationships (NR) depending on β and γ (financial glossary).

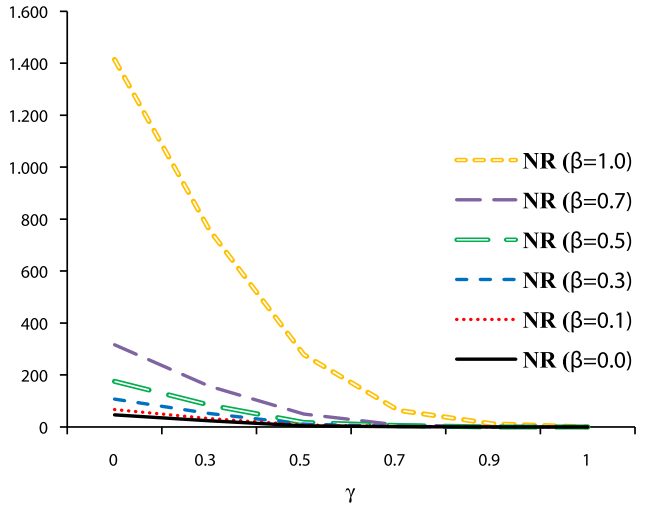

Fig. 13. Number of proposed relationships (NR) depending on β and γ (ASFA thesaurus).

VAMortgageInsurance, *InsuranceClaimForm*, *MedicalInsuranceClaimForm* are suggested.

From Table 6 it is possible to notice that for the term *Aquatic organism* from ASFA thesaurus, we get the following list of the ranked related Cyc concepts:

- AquaticOrganism.
- OrganismTypeByHabitat.

Furthermore, we get a suggested equivalently related Cyc concept *AquaticOrganism* and hierarchically related Cyc concept *OrganismTypeByHabitat*.

Fig. 14 displays two illustrative examples of Cyc KB extension with user interaction, one in fisheries & aquatic domain (with term *Rare earths* from ASFA thesaurus) and the other in financial domain (term *Recession* from financial glossary).

As proposed in our methodology, the user gets a ranked list of relevant Cyc concepts for each glossary term and confirms the relationships between the glossary term and the proposed concepts. ASFA thesaurus defines *Rare earths* as a narrow term for *Metal*. Using our methodology, for *Rare earth* the user obtains a related concept *Metal* and suggested relationship: *"Rare earth is a subclass of Metal"*. Formula (3) can be applied for an extension of the ontology with new concept *Rare earths*, which is hierarchically related to the existing ontology concept *Metal*.

For financial glossary term *Recession* the user obtains two related Cyc concepts – top ranked Recession-Economic with suggested hierarchical relationship: *"Recession equals to Recession-Economic"* and Cyc concept *Downturn* with associative relationship: *"Recession conceptually related to Downturn"*. Formula (4) defines extending the textual representation of the existing ontology concept *Recession-Economic* with lexical entries obtained from the financial glossary term *Recession*. Adding a new axiom to the ontology is specified by formula (5).

The combination of user-interaction approach with automatic concept suggestions for ontology extension prevents automatic method from establishing wrong relationships between ontology concepts, at the same time making the extension process faster and more effective than purely manual. It means that using the proposed methodology the user is able to compare Cyc concepts with glossary terms and establish relationships much faster than just using the manual search for the relevant concepts in Cyc.

Table 5
Examples of Cyc KB extension (financial glossary).

Concept	Suggested related Cyc concepts	Suggested eqv. rels.	Suggested hier.-subclass rels.	Suggested associative rels.
FOREX	**Foreign**			**Foreign**
	MoneyModeExchange		MoneyModeExchange	
	FinancialExchange		**FinancialExchange**	
	ExchangeOfUserRights			ExchangeOfUser Rights
	BondExchange			BondExchange
	MSExchangeServer			MSExchangeServer
	objectTendered			objectTendered
	NewYorkMercantile			NewYorkMercantile
	Exchange			Exchange
	MonetaryExchangeOf			MonetaryExchangeOf
	UserRights			UserRights
LIFE_INSURANCE	**LifeInsurance**	**Life**		
	Endowment-LifeInsurance	**Insurance**		**Endowment-LifeInsurance**
				FHAMortgage
	InsurancePlan		**InsurancePlan**	
	FHAMortgageInsurance			Insurance
	VAMortgageInsurance			VAMortgageInsurance
	InsuranceClaimForm			InsuranceClaimForm
	MedicalInsurance			MedicalInsurance
	ClaimForm			ClaimForm
ROTATION [Strategy]	**Movement-Rotation**		Movement-Rotation	
	IndustryOrEconomic		IndustryOrEconomic	
	SectorType		SectorType	
	Strategy		**Strategy**	
	Assets			**Assets**

Table 6
Examples of Cyc KB extension (ASFA thesaurus).

Concept	Suggested related Cyc concepts	Suggested eqv. rels.	Suggested hier.-subclass rels.	Suggested associative rels.
AQUATIC_ ORGANISMS	**AquaticOrganism**	**AquaticOrganism**		
	OrganismTypeByHabitat		**OrganismTypeByHabitat**	
SEDIMENT_MIXING	**Sediment**			**Sediment**
	Mixing		**Mixing**	
SECRETION	**SecretionEvent**			**SecretionEvent**
	Secretion-Bodily	**Secretion-Bodily**		
	Gland			Gland
	Excreting		**Excreting**	

6. Discussion

In this paper we propose a new **OntoPlus** methodology for text-driven ontology extension. The proposed **OntoPlus** methodology, presented in this paper, is based on considering advantages and shortcomings of a number of ontology learning approaches. In the presented research we are using a combination of top-down and bottom-up approaches to the ontology extension. We evaluate the proposed methodology by applying it on Cyc ontology. The top-down part involves the user identifying the keywords for extracting relevant data from the ontology, while the bottom-up part involves automatic obtaining of the relevant information available in the ontology. Usage of text mining methods involves data preprocessing, where a chain of linguistic components, such as tokenization, stop-word removal and stemming allows normalizing the textual representation of ontology concepts and a domain relevant glossary of terms with their descriptions. Text mining methods are further used for automatically determining candidate concepts in the ontology that relate to the new knowledge from the domain. A list of suggestions is provided to the user for a final decision which helps the user in narrowing down the possibilities and allows preventing the inappropriate automatic insertions into the ontology.

In contrast with many other methodologies for ontology extension, our methodology deals with ontologies and knowledge bases, potentially covering more than one domain. However, it allows restricting the area of ontology extension to a specific domain and users deal only with their sphere of interest.

OntoPlus methodology allows transforming textual information organized at different knowledge representation levels into a structured conceptualized form. Unlike the approach in [30], the proposed methodology works even if no taxonomically structured data is available on input.

Text2Onto framework [24] for ontology learning and SPRAT tool [25] for ontology population can be compared to our methodology in a number of ways. In Text2Onto, the user specifies a corpus – text collection used in ontology learning. In our case, the user defines a set of domain keywords and determines the domain relevant glossary. We expect that the efforts of keywords and glossary specification do not exceed the efforts of text corpus identification. As well as in Text2Onto, the user interaction in the proposed methodology helps to avoid adding to the ontology irrelevant concepts and relationships. Unlike the authors of Text2Onto and SPRAT tools, we do not use linguistic patterns for concepts and relations identification. Instead, we use statistically driven approaches what makes our methodology more language independent.

The experimental results show that by exploiting ontology structure information, the **OntoPlus** methodology achieves the precision of 29.5% for subclass relations identification in the financial domain and the precision of 20.3% for subclass relations iden-

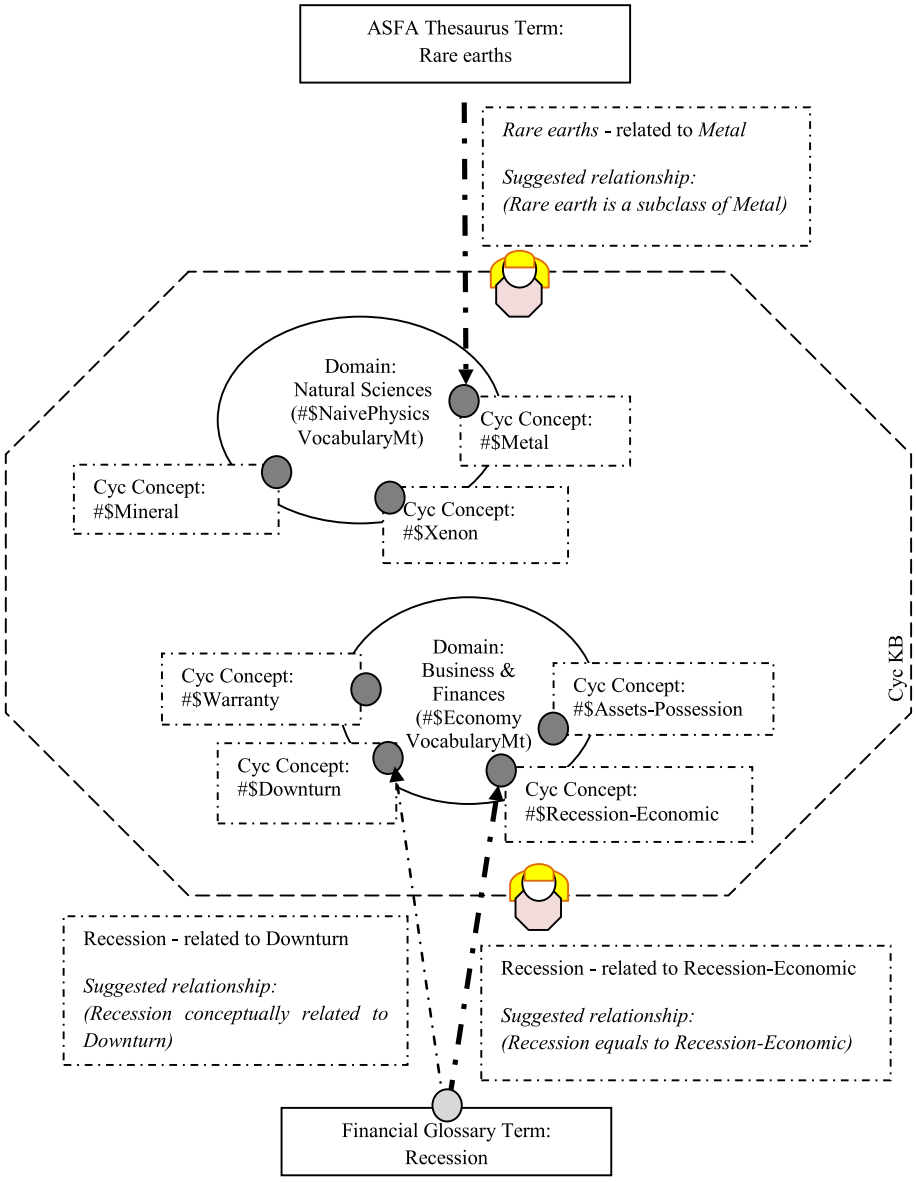

Fig. 14. Cyc KB extension – user validation.

tification in the fisheries & aquaculture domain. The precision for equivalent relations identification is 60.0% for the financial domain and 94.7% for fisheries & aquaculture domain. In comparison, the authors of Text2Onto framework obtained a precision of 17.38% for subclass-of relation identification on the subset of tourism-related texts. The creators of SPRAT tool report the precision of 48.5% for subclass identification and 48.0% for synonym recognition on a subset of Wikipedia articles about animal.

If we compare the proposed methodology to the approaches used in OntoGen [26], we can notice resemblance in using text mining technology for handling textual data and measuring similarity. While in our case, we deal with extension of general multi-domain ontology, OntoGen focuses on topic ontology construction and applies several machine learning and data visualization methods that are not used in our approach. **OntoPlus** methodology is able to perform within different domains and different information sources. For this reason, we can say that the proposed methodology goes beyond the topic ontology construction [26].

Consequently, the applicability to very large multi-domain ontologies, the possibility of diverse textual sources utilization and the usage of the language independent approaches represent the strengths of the proposed methodology. However, the proposed methodology as presented in this paper is mainly applicable for extension of the ontology, which has a sufficient lexical representation of its components.

OntoPlus methodology allows for the effective extension of very large ontologies. The methodology provides the user with required concepts and relationships in the form of the ranked list. The evaluation of **OntoPlus** methodology confirms that combining textual ontology content, ontology structure and co-occurrence information we can provide the user with higher number of concepts suitable for the ontology extension than using only concept denotations, only ontology content or only co-occurrence analysis.

The results of the experiments justify the applicability of the suggested methodology for the augmentation of large lexical ontologies such as Cyc Knowledge Base.

7. Conclusion

This paper explores the aspects of the ontology extension. **OntoPlus** methodology for text-driven ontology extension,

combining text mining methods and user-interaction approach, has been suggested and exposed to the evaluation. The evaluation of our methodology has been accomplished in two rather different domains; for the financial domain a glossary was available while for the fisheries & aquaculture domain a thesaurus has been used as a source of terms to be added to the existing ontology. Consequently, the proposed methodology works for textual data structured at different knowledge representation levels.

The main contribution of this work is the proposed methodology, where for each glossary term the user is provided with a ranked list of related ontology concepts and a ranked list of potential relations. We have found that the importance of the ontology content, structure and co-occurrence information can vary for different domains and knowledge representations used in the process of ontology extension. The best results are achieved by combining content, structure and co-occurrence information for our data in the financial domain. At the same time, ontology content and co-occurrence seem to be more important for our fisheries & aquaculture data. More exhaustive experiments across several domains involving a number of ontologies from the same domain would be needed to investigate how properties of the ontology relate to the usefulness of content, structure and co-occurrence information.

We expect our methodology to be exploited for various ontology learning and ontology alignment purposes. The future work should include further extension of Cyc Knowledge Base and using it for textual data analysis. A particular attention will be devoted to better automatic identification of hierarchical relations and extraction of different types of non-hierarchical relations.

Aknowledgements

This work was supported by the Slovenian Research Agency and the IST Programme of the EC under PASCAL2 (IST-NoE-216886) and ACTIVE (IST-2007-215040).

References

[1] L. Bradesko, L. Dali, B. Fortuna, M. Grobelnik, D. Mladenić, I. Novalija, B. Pajntar, Contextualized question answering, in: Proceedings of ITI 2010, 2010, pp. 73–78.
[2] T.R. Gruber, A translation approach to portable ontologies, Knowledge Acquisition 5 (2) (1993) 199–220.
[3] B. Chandrasekaran, J.R. Josephson, R.V. Benjamins, What are ontologies and why do we need them?, IEEE Intelligent Systems and Their Applications 14 (1999) 20–26
[4] J. Heflin, J. Hendler, Dynamic ontologies on the Web, in: Proceedings of the 17th National Conference on Artificial Intelligence, 2000, pp. 443–449.
[5] J. Curtis, G. Matthews, D. Baxter, On the effective use of Cyc in a question answering system, in: Proceedings of the IJCAI Workshop on Knowledge and Reasoning for Answering Questions (KRAQ'05), Edinburgh, Scotland, 2005, pp. 61–71.
[6] J. Curtis, J. Cabral, D. Baxter, On the application of the Cyc ontology to word sense disambiguation, in: Proceedings of the 19th International FLAIRS Conference, Melbourne Beach, FL, 2006, pp. 652–657.
[7] L. Moss, D. Sleeman, M. Sim, M. Booth, M. Daniel, L. Donaldson, C. Gilhooly, M. Hughes, J. Kinsella, Ontology-driven hypothesis generation to explain anomalous patient responses to treatment, Knowledge-Based Systems 23 (4) (2010) 309–315.
[8] D. Sánchez, M. Batet, A. Valls, K. Gibert, Ontology-driven web-based semantic similarity, Journal of Intelligent Information Systems 35 (3) (2009) 383–413.
[9] Cycorp, Inc., What's in Cyc. <http://www.cyc.com/cyc/technology/whatiscyc_dir/whatsincyc>.
[10] I. Novalija, D. Mladenić, Extending ontologies for annotating business news, in: Proceedings of SiKDD 2008, 2008, pp. 186–190.
[11] N.F. Noy, C. Hafner, The state of the art in ontology design: a survey and comparative review, Artificial Intelligence Magazine (1997) 53–74.
[12] Cycorp, Inc. <http://www.cyc.com>.
[13] C.R. Harvey, Yahoo Financial Glossary, Fuqua School of Business, Duke University, 2003.
[14] ASFA Thesaurus. <http://www.4.fao.org/asfa/asfa.htm>.
[15] P. Buitelaar, P. Cimiano, B. Magnini (Eds.), Ontology Learning from Text: Methods, Evaluation and Applications, IOS Press, 2005.
[16] M.-L. Reinberger, P. Spyns, Unsupervised text mining for the learning of DOGMA-inspired ontologies, in: P. Buitelaar, S. Handschuh, B. Magnini (Eds.), Ontology Learning from Text: Methods, Evaluation and Applications, Springer, 2005.
[17] M. Grobelnik, D. Mladenic, Knowledge discovery for ontology construction, in: J. Davies, R. Studer, P. Warren (Eds.), Semantic Web Technologies: Trends and Research in Ontology-Based Systems, John Wiley & Sons, 2006, pp. 9–27.
[18] F. Burkhardt, J.A. Gulla, J. Liu, C. Weiss, J. Zhou, Semi automatic ontology engineering in business applications, in: Workshop Applications of Semantic Technologies, INFORMATIK 2008, 2008.
[19] T. Sabrina, A. Rosni, T. Enyakong, Extending ontology tree using NLP technique, in: Proceedings of National Conference on Research & Development in Computer Science REDECS 2001, 2001.
[20] P. Cimiano, A. Hotho, S. Staab, Learning concept hierarchies from text corpora using formal concept analysis, Journal of Artificial Intelligence Research (JAIR) 24 (2005) 305–339.
[21] E. Agirre, O. Ansa, E. Hovy, D. Martínez, Enriching very large ontologies using the Www, in: Proceedings of ECAI 2000, Workshop on Ontology Learning, 2000.
[22] WordNet – Princeton University Cognitive Science Laboratory. <http://wordnet.princeton.edu>.
[23] R. Prieto-Diaz, A faceted approach to building ontologies, in: Proceedings of the 2003 IEEE International Conference on Information Reuse and Integration, 2003, pp. 458–465.
[24] P. Cimiano, J. Völker, Text2Onto a framework for ontology learning and data-driven change discovery, in: Proceedings of NLDB 2005, 2005, pp. 227–238.
[25] D. Maynard, A. Funk, W. Peters, SPRAT: a tool for automatic semantic pattern-based ontology population, in: Proceedings of International Conference for Digital Libraries and the Semantic Web, Trento, Italy, 2009.
[26] B. Fortuna, M. Grobelnik, D. Mladenić, OntoGen: semi-automatic ontology editor, HCI 9 (2007) 309–318.
[27] P. Shah, D. Schneider, C. Matuszek, R.C. Kahlert, B. Aldag, D. Baxter, J. Cabral, M. Witbrock, J. Curtis, Automated population of Cyc: extracting information about named-entities from the web, in: Proceedings of the 19th International FLAIRS Conference, 2006, pp. 153–158.
[28] M. Witbrock, D. Baxter, J. Curtis, D. Schneider, R. Kahlert, P. Miraglia, P. Wagner, K. Panton, G. Matthews, A. Vizedom, An interactive dialogue system for knowledge acquisition in Cyc, in: Proceedings of the Workshop on Mixed-Initiative Intelligent Systems, 2003, pp. 138–145.
[29] O. Medelyan, C. Legg, Integrating Cyc and Wikipedia: folksonomy meets rigorously defined common-sense, in: Proceedings of Wiki-AI Workshop at the AAAI'08 Conference, Chicago, US, 2008.
[30] S. Sarjant, C. Legg, M. Robinson, O. Medelyan, "All you can eat" ontology-building: feeding Wikipedia to Cyc, in: Proceedings of the 2009 IEEE/WIC/ACM International Conference on Web Intelligence, WI'09, Milan, Italy, 2009, pp. 341–348.
[31] P. Cimiano, A. Pivk, L. Schmidt-Thieme, S. Staab, Learning taxonomic relations from heterogeneous evidence, in: Proceedings of ECAI 2004, Workshop on Ontology Learning and Population, 2004.
[32] A. Maedche, S. Staab, Discovering conceptual relations from text, in: W. Horn (Ed.), ECAI 2000, Proceedings of the 14th European Conference on Artificial Intelligence, Berlin, August 21–25, 2000, IOS Press, Amsterdam, 2000, pp. 321–324.
[33] G. Heyer, M. Läuter, U. Quasthoff, T. Wittig, C. Wolff, Learning Relations using collocations, in: Proceedings of IJCAI-2001, Workshop on Ontology Learning, 2001, pp. 19–24.
[34] P.D. Turney, Mining the Web for synonyms: PMI-IR versus LSA on TOEFL, in: Proceedings of the 12th European Conference on Machine Learning, 2001, pp. 491–502.
[35] T. Štajner, D. Mladenić, Entity resolution in texts using statistical learning and ontologies, in: Proceedings of Asian Semantic Web Conference, 2009, pp. 91–104.
[36] A. Maedche, S. Staab, Ontology learning for the Semantic Web, IEEE Intelligent Systems 16 2 (2001) 72–79.
[37] The Syntax of CycL. <http://www.cyc.com/cycdoc/ref/cycl-syntax.html>.
[38] Contexts in Cyc. <http://www.cyc.com/cycdoc/course/contexts-basic-module.html>.
[39] K. Dellschaft, S. Staab, Strategies for the evaluation of ontology learning, in: P. Cimiano, P. Buitelaar (Eds.), Bridging the Gap between Text and Knowledge – Selected Contributions to Ontology Learning and Population from Text, IOS Press, 2008, pp. 253–272.
[40] A. Maedche, S. Staab, Measuring similarity between ontologies, in: Proceedings of the European Conference on Knowledge Acquisition and Management – EKAW-2002, Madrid, Spain, 2002, pp. 251–263.
[41] U. Hahn, K. Schnattinger, Towards text knowledge engineering, in: Proceedings of the AAAI'98, 1998, pp. 129–144.
[42] M. Deshpande, G. Karypis, Item-based top-*N* recommendation algorithms, ACM Transactions on Information Systems 22 (1) (2004) 143–177.
[43] Yahoo! Finance. <http://finance.yahoo.com>.
[44] I.V. Levenshtein, Binary codes capable of correcting deletions, insertions and reversals, Cybernetics and Control Theory 10 (8) (1966) 707–710.