View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Research Repository and Portal - University of the West of Scotland

UWS Academic Portal

Taxo-Semantics

Angermann, Heiko; Pervez, Zeeshan; Ramzan, Naeem

Published in:
Decision Support Systems

DOI:
10.1016/j.dss.2017.04.001

Published: 01/06/2017

Document Version
Peer reviewed version

Link to publication on the UWS Academic Portal

Citation for published version (APA):
Angermann, H., Pervez, Z., & Ramzan, N. (2017). Taxo-Semantics: Assessing similarity between multi-word expressions for extending e-catalogs. Decision Support Systems, 98, 10-25. https://doi.org/10.1016/j.dss.2017.04.001

General rights
Copyright and moral rights for the publications made accessible in the UWS Academic Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
If you believe that this document breaches copyright please contact pure@uws.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 17 Sep 2019

Taxo-Semantics: Assessing similarity between multi-word expressions for extending e-catalogs

Heiko Angermann[a,*], Zeeshan Pervez[a], Naeem Ramzan[a]

[a] *University of the West of Scotland, School of Engineering and Computing, High Street, Paisley PA1 2BE, United Kingdom*

Abstract

Taxonomies, also named directories, are utilized in e-catalogs to classify goods in a hierarchical manner with the help of concepts. If there is a need to create new concepts when modifying the taxonomy, the semantic similarity between the provided concepts has to be assessed properly. Existing semantic similarity assessment techniques lack in a comprehensive support for e-commerce, as those are not supporting multi-word expressions, multilingualism, the import/export to relational databases, and supervised user-involvement. This paper proposes *Taxo-Semantics*, a decision support system that is based on the progress in taxonomy matching to match each expression against various sources of background knowledge. The similarity assessment is based on providing three different matching strategies: a lexical-based strategy named *Taxo-Semantics-Label*, the strategy *Taxo-Semantics-Bk*, which is using different sources of background knowledge, and the strategy *Taxo-Semantics-User* that is providing user-involvement. The proposed system includes a translating service to analyze non-English concepts with the help of the WordNet lexicon, can parse taxonomies of relational databases, supports user-involvement to match single sequences with WordNet, and is capable to analyze each sequence as (sub)-taxonomy. The three proposed matching strategies significantly outperformed existing techniques. *Taxo-Semantics-Label* could improve the accuracy result by more than 7 % as compared to state-of-the-art lexical techniques. *Taxo-Semantics-Bk* could improve the accuracy compared to structure-based techniques by more than 8 %. And, *Taxo-Semantics-User* could additionally increase the accuracy by on average 23 %.

Keywords: Concept Similarity, Electronic Commerce, Electronic Catalog, Logic Programming

*Corresponding author: Heiko Angermann, University of the West of Scotland,
School of Engineering and Computing, High Street, Paisley PA1 2BE, United Kingdom, Tel +44 (0) 141 848 3648
Email addresses: B00274523@studentmail.uws.ac.uk (Heiko Angermann), zeeshan.pervez@uws.ac.uk (Zeeshan Pervez), naeem.ramzan@uws.ac.uk (Naeem Ramzan)

1. Introduction

Taxonomies are subcategories of ontologies, using hierarchically ordered concepts to model a field of interest in a formal way (Sanchez and Batet (2013)). While keyword search is known as a quick solution for finding specific products, a hierarchical representation of a domain has its merits for navigation, and for exploring similar items (Angermann and Ramzan (2016a)). The creation of the taxonomy is either performed through expert(s) knowing the domain in detail, by extracting the taxonomy from a text corpus, or by referring to a standard taxonomy (Meijer et al. (2014)) (e.g. *Global Language of Business* (GS1), the *United Nations Standard Products and Services Code* (UNSPSC), or the *Classification and Product Description* (eCl@ss)). Such standard taxonomies, as well as custom e-commerce taxonomies often use Multi-Word Expressions (MWEs) as labels for the concepts combining semantically less or more similar sequences as a label.

In the age of the digital customer, it is now possible to derive the customers' preferences with the concepts (Sharma and Dey (2015); Angermann and Ramzan (2016b)). The preferences can be used for providing personalized directories (e.g. in Angermann and Ramzan (2016c)), but new concepts have to be created accordingly, as the semantics of the taxonomy is changing. The new concepts can again be created by the expert, or through systems performing a *Semantic Similarity Assessment* (SSA) measure. The lexicon-based SSA measures are taking into account literal values, respectively the terminological heterogeneity existing between concepts (Chuang and Chien (2003); Furlan et al. (2013)). In contrast, the structure-based SSA techniques are detecting similarity by the help of the taxonomy-based *is-A* hierarchy (Ahsae et al. (2012)). Hereby, the conceptual heterogeneity is used to derive the path-length distance between the two concepts based on the depth of the concepts inside the taxonomy. The information-based SSA measures are also using the conceptual heterogeneity, but are exploiting the additional content assigned to the concepts and their probability to occur inside the taxonomy. For inferring further similarities between the concepts, the semantic lexicon WordNet is widely used (Miller (1995)). The authors in Ma et al. (2013) are manually mapping the concepts of the source taxonomy to the synsets in WordNet before applying a lexical-based measure. Another approach in Kim et al. (2013) is determining the path of the concepts in WordNet, before computing structure-based similarity. The approach in Nguyen and Conrad (2013) analyzes the indirect connections between the WordNet entities.

However, as recent approaches do not consider the comparison between MWEs in e-commerce, the need for creating new concepts, the output of the assessment result in relational database format, or the support of multilingualism as provided in other domains (e.g. Hogenboom et al. (2013)), the proposed systems are not comprehensive enough to be used in the e-catalog/e-commerce domain.

To provide a system for comparing similarity between concepts in e-catalogs, and to use the comparison results for extending the taxonomy, the decision support system *Taxo-Semantics* is presented. The proposed system is build upon the progress made in the related research field *Taxonomy Matching* (TM) and differs from existing techniques in five ways. Firstly, non-English taxonomies can not be analyzed with the help of WordNet when a translator is missing, as WordNet only contains English synsets. To overcome multilingualism, a dictionary can be included in *Taxo-Semantics* to match non-English sequences. Secondly, for parsing real-world e-commerce taxonomies, the approach must be capable to process taxonomies captured in *Structured Query Language* (SQL). In *Taxo-Semantics*, the SQL import/export format *Comma Separated Values* (CSV) can be parsed to analyze and extend taxonomies coming from relational databases. Thirdly, although the latest *Ontology Alignment Evaluation Initiative* (OAEI)[1] campaigns evident that user-involvement can significantly affect the assessment quality results (Cheatham et al. (2016)), either no user-involvement is supported in recent systems, or the provided user-involvement is limited to identify the most-related WordNet synset. In *Taxo-Semantics*, the user-involvement is supported in a supervised way. Through this, it helps non-expert users, as well as expert users to detect the related WordNet synset for each MWE seqeunce. In addition, the user-involvement is used as further technique to affect the SSA. Fourthly, recent approaches are focussing on single labels, although, the most real-world e-commerce applications like Amazon[2] are using MWEs. To compare concepts using MWEs, *Taxo-Semantics* is providing three different and scalable matching strategies. And fifthly, recent approaches do not consider the usage of the assessment result for creating new concepts. *Taxo-Semantics* is capable to use the assessment result for creating mediator concepts that generalize the semantically similar concepts. In the end, *Taxo-Semantics* provides the following contributions to the field of decision support systems:

- A decision support system that helps (non)-experts to semantically compare concepts of tax-

[1] http://oaei.ontologymatching.org/
[2] http://www.amazon.com

onomies used in e-commerce.

- A decision support system that performs the assessment process as matching operation.
- A decision support system that involves the user during the matching/assessment process.
- A decision support system, which makes use of the matching result to extend the taxonomy.

The remainder of the paper is organized as follows. In Section 2, the problems of SSA, and TM are formulated, as well as the usage of MWEs in e-catalogs. The decision support system *TaxoSemantic* is presented in Section 3. In Section 4, the system is evaluated with the help of three e-catalogs. The work concludes in Section 5.

2. Problem Formulation

Formally, a **Taxonomy** T is an out-tree, see Equation 1:

$$T = \{C, E\}, \tag{1}$$

which is using a set of concepts C for describing terms with the help of a label, and a set of edges E connecting less general with more general concepts. The less general concepts are formally named **Sub Concept** of its super-ordinated concept. In the example shown in Figure 1, the concept "Car Accessoires & Parts" *is-A* sub concept of "Car & Motorbike", which *is-A* sub concept of "Shop by Department". The super-ordinated concept is named **Super Concept**. Consequently, "Shop by Department" *is-A* super concept of "Car & Motorbike", which *is-A* super concept of "Car Accessoires & Parts", as well as of "Tools & Equipment", "Sat Nav & Car Electronics", and "Motorbike Accessoires & Parts". Concepts sharing the same super concept are referred as **Sibling Concepts**. A **Root Concept** has no more generalized super concept. In the illustrated taxonomy, "Shop by Department" is the root concept of the e-catalog. Optional, each concept can have additional information like a description (e.g. "tools for reparing a motorbike"), and a set of properties (e.g. "color", "power consumption", "size").

In recent e-commerce applications, the labels mainly consist of a combination of different word sequences, named **Multi-Word Expressions** (MWEs). MWEs are defined as idiosyncratic interpretations that cross word boundaries (Sag et al. (2002)). For example, the Amazon concept "Car

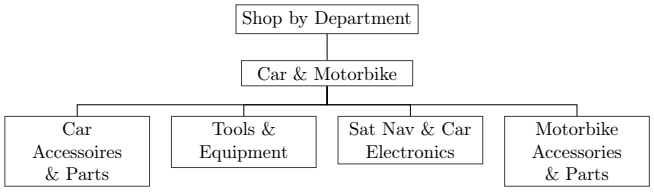


Fig. 1. A subset of the Amazon product taxonomy.

& Motorbike", includes four less generalized concepts that all consist of MWEs. Or, the Walmart concept "Auto Detailing & Car Care" includes five concepts, which are throughout labelled with MWEs, see Figure 2. However, for the most widely used background resource for inferring semantic similarity WordNet, only 41 % of the synsets are MWEs (Sag et al. (2002)).

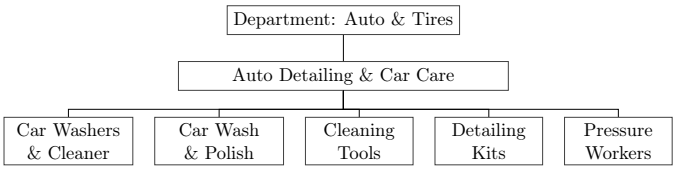


Fig. 2. A subset of the Walmart product taxonomy.

Semantic Similarity Assessment (SSA) is the task of finding the cognitive and methodical homogeneity between a pair of concepts, formally sim_{C_1,C_2}. For example, between the concepts "Car Accessoires & Parts", and "Tools & Equipment", see Figure 3. Two terms are semantically similar, if their meanings are close, or the concepts are sharing common attributes (Li et al. (2013)). **Taxonomy Matching** has a similar aim, namely to find the correspondences between concepts of two taxonomies. The correspondences are computed during a matching operation (Peukert et al. (2012)), which includes a matching strategy that defines how the similarities should be assessed. Current matching systems combine different techniques to increase the assessment accuracy (Shvaiko and Euzenat (2013)). In contrast to SSA, such approaches are annually evaluated by the OAEI.

3. Proposed system Taxo-Semantics

This section explains the system *Taxo-Semantics*. The explanation starts by detailing the methodology used to detect similarity between concepts. Afterwards, its implementation in logic programming language Prolog is presented.

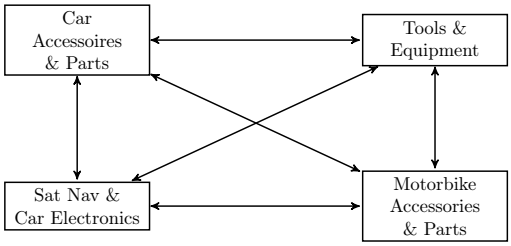


Fig. 3. Semantic similarity assessment between concepts of the Amazon taxonomy.

3.1. Taxo-Semantics method

As the hierarchical structure inside the taxonomy already provides different concept types, this relationships can be used to create concept pairs. In *Taxo-Semantics*, the sibling sub concepts are taken into account as concept pairs. Thus, each concept detailing a super concept is compared with every sibling concept. Because of the by nature flexible number of concepts, each super concept can have N pairs, see Equation 2:

$$N = n!/((n-k)! * k), \tag{2}$$

where n is the number of sibling sub concepts, and k is the number of concepts being part of a concept pair. In *Taxo-Semantics*, each pair consists of two concepts. For example, the super concept "Car & Motorbike" given in Figure 1 would create six pairs. The sub concept "Car Accessoires & Parts" would be compared with "Tools & Equipment", "Sat Nav & Car Electronics", and "Motorbike Accessories & Parts". Its sibling concept "Tools & Equipment" would be compared with "Sat Nav & Car Electronics", as well as "Motorbike Accessories & Parts". Finally, the sub concept "Sat Nav & Car Electronics" would be compared with "Motorbike Accessories & Parts". Through this, each sibling concept is compared with each other sibling concept. By default, *Taxo-Semantics* is considered to compare the sub concepts belonging to the identical super concept. However, a comparison between all sub concepts of a taxonomy, no matter to which super concept those belong, would also be possible. Than, the root concept has to be considered as super-ordinated concept, instead of using the super concept. This is necessary if redundant sub concepts exist and the assessment should be performed overall concepts.

For assessing the similarity between a pair, *Taxo-Semantics* does not use a static assessment

procedure. *Taxo-Semantics* is based on the progress in the field of taxonomy matching, and performs similar to a matching system, e.g. in Faria et al. (2014); Jimenez-Ruiz et al. (2014). Such matching systems usually provide different matching techniques inside a library to be used in a flexible matching strategy according to the users' selection. Through this, a matching strategy can, in contrast to a static assessment procedure react on the characteristics of the domain, i.e. the quality of the taxonomy. For example, if the differences between the concepts is mainly based on the literal values, the matching strategy should focus on string- or lexical-based techniques. Or, if the differences between the concepts is mainly based on conceptual values, the matching strategy should focus on structure-based techniques. This flexibility is considered in *Taxo-Semantics* by providing three different matching strategies, which in turn can use techniques of other matching strategies: *Taxo-Semantics-Label*, in the following shortened as TS-Label, *Taxo-Semantics-Bk*, in the following shortened as TS-Bk, and *Taxo-Semantics-User*, in the following shortened as TS-User. TS-Label is comparing the labels of the concepts with a lexical-based technique. TS-Bk is based on background-knowledge. To do so, each meaningful sequence of the MWEs is searched in WordNet. After that, for each related synset, the (sub)-taxonomy is created based on its super concepts in WordNet, named hypernym. The concepts being included in the (sub)-taxonomy are compared using the shortest path distance existing between the two concepts (Lingling et al. (2013)). Hereby the path distance is based on the depth of the less general synset for the two concepts, not on the complete WordNet taxonomy. In addition, this strategy can combine the lexical-based technique of the former strategy, as well as a language-based technique performing on the WordNet gloss, and a content-based technique using the properties of the concepts. TS-User is based on TS-Bk, but additionally supports user-involvement. Hereby, the user can state if a similarity between the concepts exists or not, which is afterwards combined with the result of the other used techniques. For all strategies, the user can define a dynamic threshold. If the value resulted by the system is equal or higher than the threshold, the concepts inside the pair are considered as similar.

3.1.1. Background Sources

Background sources are used to help inferring further relationships between concepts and thus help assessing similarity between concepts/taxonomies (Shvaiko and Euzenat (2013)). The latest OAEI campaigns evident that the taxonomy matching systems perform better, than more resources of background knowledge they are using (Cheatham et al. (2016)). In *Taxo-Semantics*, four different

sources of background knowledge are used: the semantic lexicon WordNet, a dictionary containing all English root words, a list containing punctmarks, as well as a thesaurus.

The semantic lexicon WordNet, its synsets and hypernyms, are used for performing a structure-based analyzes between the pair. A hypernym, in the form of: $Hypernym = (Synset, Synset)$ specifies that the second argument is a super-ordinated synset of the first synset, see Figure 4. Thus, it represents the relationships analogues to the *is-A* hierarchy. In addition, *Taxo-Semantics* is capable to make use of the gloss that is assigned to every WordNet synset, in the form of: $Gloss = (Synset, Text)$. A gloss gives a brief definition of the synset and, in most cases, one or more short sentences illustrating the use of synset members (Fellbaum (1998)). With the help of the gloss, a language-based comparison on a larger text corpus can be performed, also if the taxonomy is initially not supported with concept descriptions, as given in the most e-commerce applications. The *Summer Institute of Linguistics dictionary*[3] is a resource that contains all English root words. In *Taxo-Semantics*, it is used for supporting the before-mentioned language-based technique. Punctmarks are also considered as irrelevant seqeunces having no semantically rich influence on the label of a concept, as well as on the gloss. A thesaurus can be used to match between the input language, and between English. This is necessary as WordNet only contains English synsets. In *Taxo-Semantics*, the Google translator is used to result a list containing the initial sequence, as well as the sequence expressed in English (for further details see Aiken et al. (2009)).

3.1.2. Matching Operation

Taxo-Semantics is performing the matching operation firstly between the concepts and the background knowledge, and secondly, between the sibling concepts. The workflow to assess the similarity consists of seven steps.

1. **Select Matching Strategy** defines the strategy to be used for performing the match, as well as the threshold to state if a similarity exists or not. As explained above, a user can choose between three matching strategies. The complete strategy is defined as a six-tuple, see Equation 3:

$$ST = \{STR, THH, BKL, BKG, BKP, UIV\}, \tag{3}$$

[3] http://www.sil.org/

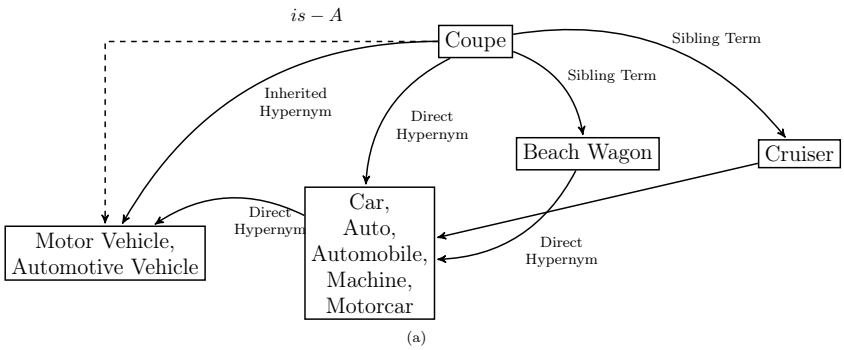


Fig. 4. An exemplary result for querying "Coupe" in WordNet: the result represented as a graph including an inferred *is-A* relationship.

where STR defines the strategy, THH the threshold, and UIV if user-involvement should be supported or not. The variables BKL (lexical-based), BKG (language-based), and BKP (content-based) define if the matching strategy TS-Bk should combine the structure-based technique with other techniques.

2. **Input Source Taxonomy** asks the user for choosing the input taxonomy, and to define the natural language of the taxonomy. The concepts of the input taxonomy are parsed according to its file format. Hereby, the concepts are investigated against its super and sub concepts, which results that a concept having no super ordinated concept is defined as the root concept, and the concepts having a super-ordinated concept are sub concepts. In the case of the concept includes properties, those are added accordingly. Through this, each concept is defined in a four-tuple, see Equation 4:

$$TA = \{SIS, SNS, SIP, [PRO_1, \ldots, PRO_N]\}, \tag{4}$$

where SIS is the identifier of the concept, SNS is the label of the concept, SIP is the identifier of the super-ordinated concept, and PRO includes the properties. Through this, *Taxo-Semantics* is not limited to a specific number of levels, because a super concept can of course again be a sub concept of a more general concept. On the base of the defined concepts, the concept pairs are created as a three-tuple, see Equation 5:

$$CP = \{SIS1, SIS2, SIP\}, \tag{5}$$

where $SIS1$ and $SIS2$ are two sibling concepts of SIP. If the language of the taxonomy is not English, the concepts are translated using a thesaurus in the form of a three-tuple, see Equation 6:

$$TR = \{LAN, SNS, SNL\}, \tag{6}$$

where LAN is defining the language, SNS is the label expressed in the initial language, and SNL is the label in English.

3. **Preprocess Background Knowledge** loads the necessary background resources if the strategy TS-Bk or TS-User is choosen. As explained above, the synsets, hypernyms, and gloss of WordNet are used as main background resources. Furthermore, the dictionary and the punctation marks (semi-colon, comma, full stop, colon, quotation mark, question mark, exclamation mark, brackets, hypen, dashes, apostrophe, braces, slash) are used for supporting the language-based technique, and to help detecting the most-related WordNet synsets. The system establishes for each concept one or multiple related WordNet synset(s). To do so, *Taxo-Semantics* reduces the label to its root word through removing stop words and abbreviations that are not found in the dictionary. For example, the label "Pendulum Jigsaw CARVEX PS 420" is reduced to "pendulum jigsaw", because the provider specific sequences "CARVEX", "PS", and "420" won't be found within WordNet. Note that the abbreviation and number are only reduced for matching with WordNet, for the lexical-based technique, the abbreviations and numbers remains. This is important if the main difference between concepts is mainly affected by a company-specific number (e.g. "Boeing 737" and "Boeing 777"). After the reduction to the root word, the single root word, and the root word consisting of MWEs, are searched in WordNet synsets as follows:
   (a) Firstly, the complete root word is searched in WordNet, as WordNet also contains a small portion of MWEs.
   (b) Secondly, if no related synset is found, the single sequences are queried. For example, the term "pendulum jigsaw" can not be found as WordNet synset. However, there is a synset for "pendulum", and another synset for "jigsaw".
   (c) Thirdly, if the first and second search fails, the sequence is assigned to the most general subsumer in WordNet. Alternatively, the single sequence of the MWE can be ignored.

   For each detection, the synset is displayed to the user along with its associated gloss. For example, the sequence "car" can be found several times in WordNet. Each synset containg "car"

as word of speech is displayed to the user along with its synonyms (e.g. “car”, “automobile”, etc.), and the gloss (e.g. “a motor vehicle with four wheels; usually propelled by an internal combustion engine”). Through this, the user can choose the most-related WordNet synset for each sequence. After defining the synset(s), each concept is assigned accordingly with one or multiple two-tuples, see Equation 7:

$$CS = \{SIS, SIY\}, \tag{7}$$

where SIY is one of the related WordNet synset(s) for a concept. Afterwards, *Taxo-Semantics* creates (sub)-taxonomies for each verified synset. The (sub)-taxonomies are having the direct hypernym of the detected synset, and the inherited hypernyms as more general concepts. Consequently, the most general concept is the WordNet root concept.

4. **Perform Similarity Assessment** is computing the semantic similarity according to the choosen strategy. TS-Label is comparing the initial labels with utilizing the ISub[4] library. The lexical-based comparison result is captured inside a four-tuple, see Equation 8:

$$IS = \{SIS1, SIS2, SIP, SIS\}, \tag{8}$$

where SIS is the similarity result. TS-Bk starts by comparing the (sub)-taxonomies using the shortest path measure, a structure-based variant of the technique provided in Rada et al. (1989), see Equation 9:

$$TAS = 2 * deep_{max} - len(SIY1, SIY2), \tag{9}$$

where $len(SIY1, SIY2)$ is the length existing between two synsets $SIY1$ and $SIY2$, and $deep_{max}$ is the depth inside WordNet for the synset appearing deeper inside WordNet. For resulting the maximum length, a search for the first common subsumer is performed. If the common subsumer is very close to both synsets, both synsets are considered as similar. And, if the subsumer is not very close to both synsets, the synsets are considered as not similar, because of a longer path distance. This comparison result is captured inside a four-tuple, see

[4]http://www.swi-prolog.org/

Equation 9:

$$SE = \{SIS1, SIS2, SIP, TAS\}, \tag{10}$$

where TAS expresses the path closeness. If TS-Bk is additionally using the language-based analysis, the glosses are compared, after using elimination and tokenization. Again, its comparison result is stored inside a four-tuple, see Equation 11:

$$JA = \{SIS1, SIS2, SIP, JAS\}, \tag{11}$$

where JAS is the Jaccard similarity existing between two WordNet glosses. If TS-Bk should be combined with the lexical-based technique, the algorithm of TS-Label is performed. And finally, if a content-based technique is desired, the properties of the concepts are compared using Jaccard similarity. Its comparison result between the two sets is captured inside a four-tuple, see Equation 12:

$$PR = \{SIS1, SIS2, SIP, PAS\}, \tag{12}$$

where PAS is the Jaccard similarity result between two compared lists of properties. If the matching is supported using user-involvement, the user can additionally define if a semantic similarity exists between the two concepts or not. The user has to assess the similarity by the help of both initial labels. The assessment is captured in a four-tuple, see Equation 13:

$$UI = \{SIS1, SIS2, SIP, UIV\}, \tag{13}$$

where UIV is the rating of the user, which can be zero ("0") for stating not similar, or one ("1") for stating that both labels, respectively the concepts, are similar.

5. **Combine Strategy Results** is necessary if the similarity assessment is performed using the strategy TS-Bk, and the strategy is using a combination of multiple techniques. Consequently, it calculates the average mean of the different measures. Hereby, each technique is considered with the same weight.
6. **Filter Irrelevant Similarities** is comparing the final similarity result with the threshold to state if both concepts are semantically similar or not.

7. **Output Assessed Pairs** exports the final result in CSV data format, more precisely, the included five-tuples, see Equation 14:

$$S = \{SIP, SIS1, SIS2, THH, SAS\}, \tag{14}$$

where SAS identifies, if both concepts are semantically similar ("yes"), or not ("no").

3.1.3. Extending Taxonomies

Taxo-Semantics can use the output of the assessment to help the user defining more complex relationships between sub-ordinated and super-ordinated concepts, instead of connecting those with a common super concept. In the following, this type of concept is referred as mediator concept. Formally, a **Mediator Concept** is a sub-ordinated concept of a super concept, and super-ordinates N sub-ordinated concepts of the super concept. For example, the concepts "Car Washers & Cleaner" and "Car Wash & Polish" can have the mediator concept "Car Washing & Cleaning Material", or the concepts "Cleaning Tools" and "Pressure Workers" can have the mediator concept "Cleaning and Pressure Equipment", see Figure 5. This presupposes that the sub concepts of a mediator concept are detected as semantically similar, and the sub concepts that are not part of the mediator concept, are detected as not similar. For example, the concept "Detailing Kits" has no similar sibling concept, thats why no mediator concept is created. Creating such mediator concepts is required when there is a need to detail the taxonomy, for example when providing personalised directories. The creation of mediator concepts makes use of the user, the assessment result, and performs in two steps:

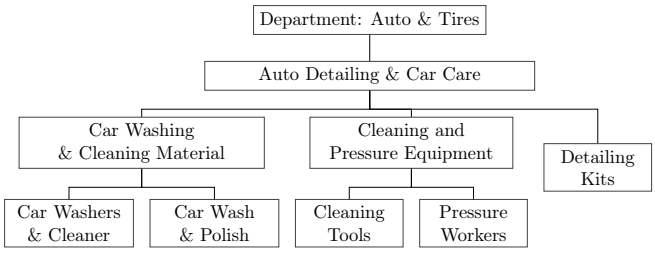


Fig. 5. A subset of the exemplary extended Walmart product taxonomy.

1. For each concept, the label is displayed to the user. According to the results of the similarity comparison, the semantically similar sibling concepts are detected. Each label is displayed

to the user to define a label for the mediator concept. Finally, the mediator is added to the knowledge base as a four-tuple, see Equation 15

$$M = \{SIS, NBS, SIP, MED\}, \tag{15}$$

where NBS is a list containing all semantically similar sibling concepts, and MED is the label of the mediator concept.

2. Similar to the last step of the matching operation, the mediator concepts are output in CSV.

3.2. Taxo-Semantics implementation

To illustrate *Taxo-Semantics*, the logic programming language Prolog is used. In Prolog, a knowledge base consisting of facts, rules, and queries is used to structure the e-catalog and the system. Facts are predicates, which in contrast to rules, do not query other predicates. Each predicate (e.g. $main(ID, Name)$) includes a functor, i.e. the name of the predicate, and a set of arguments captured between brackets (e.g. $(ID, Name)$). The short form to represent predicates is to write the number of arguments behind the functor (e.g. $main/2$). The arguments can be variables (starting with a capital letter), lists, atoms, or strings. A list (written as $[\dots]$) can include multiple atoms, strings, or other lists. Through its data structure, logic approaches are geared to deal efficiently with larger sets of concepts, but also with small business taxonomies, and provide an effective approach for knowledge management in e-commerce, especially for taxonomical engineering (Gomez-Perez et al. (2006)). The final query in *Taxo-Semantics* ($assess/0$) performs with the help of seven foregone rules, see Listing 1.

```
assess:-
  useroutput('Define if assessment or extension (0/1)'),userinput(EXT),
  (EXT = 0,input,strategy,preprocess,matching,combine,filter,output),
  (EXT = 1,input,strategy,preprocess,matching,combine,filter,output,extend).
```

Listing 1: s/0

The input of the taxonomy is done using rule $input/0$, see Listing 2.

```
input:-
  inputdefine(POS,PAT,LAN),(inputprolog(POS,PAT,LAN);inputcsv(POS,PAT,LAN)).

inputdefine(POS,PAT,LAN):-
  userinput(PAT),datamodel(PAT,POS),writeln('Define language'),userinput(LAN).

inputprolog("pl",PAT,LAN):-
  loadprolog(PAT),identifypairs(PAR),createpairs(PAR),translate(LAN).

inputcsv("csv",PAT,LAN):-
  loadcsv(PAT),forall(ta(SIS,SNS,SIP,PRP),(assert(ta(SIS,SNS,SIP,PRP)),
  retract(ta(SIS,SNS,SIP,PRP)))),identifypairs(PAR),createpairs(PAR),translate(LAN).

inputtranslate(LAN):-
  (LAN = eng,true,!);(not(LAN = eng),useroutput('Define a CSV2 dictionary'),
  userinput(PAT),loadcsv(PAT),forall(ta(ID,SNS,_,_),(tr(LAN,SNS,SNL),
  retract(ta(ID,SNS,_,_)),assert(ta(ID,SNL,_,_))))).
```

Listing 2: si/0

The used rule $inputdefine/3$ is taken to define the taxonomy (PAT), its data model (POS), and the natural language of the taxonomy (LAN). Another rule $inputprolog/3$ is using $identifypairs/1$ for identifying the super concepts having minimum two sub concepts. For each super concept, the concept pairs are asserted to the knowledge base ($cp/3$) using predicate $createpairs/1$. If the taxonomy is stored in CSV format, rule $inputcsv/3$ satisfies, instead of $inputprolog/3$. And finally, if the language is not English, a dictionary must be defined in $translate/1$.

The strategy, including the threshold, is queried with the rule $strategy/0$, see Listing 3. In $techniques/4$, the user can define if the proposed technique should be used (1) or not (0) to support the on background knowledge based strategy. A lexicon-based technique performing on the labels (BKL), a language-based technique performing on the WordNet gloss assigned to the synset(s) (BKG), and a content-based technique using the concepts' properties (BKP). Furthermore, the user has to define, if user-involvement should be supported or not in UIV. The final strategy is captured in $st/6$.

```
ss:-
  useroutput('Define matching strategy: TS-Label (1), or TS-Bk (2)'),userinput(STR),
  useroutput('Define a threshold (0.0-1.0)'),userinput(THH),
  (STR = 2,techniques(BKL,BKG,BKP,UIV),assert(st(STR,THH,BKL,BKG,BKP,UIV)));
  (STR = 1,assert(st(STR,THH,0,0,0,0))).
```

Listing 3: ss/0

The preprocessing loads the background resources using rule *preprocess*/0, see Listing 4. It also searches for each concept the related WordNet synset(s) to calculate the (sub)-taxonomies.

```
preprocess:-
  st(STR,_,_),((STR = 2,useroutput('Define bk directory'),
  userinput(DIR),wordnet(DIR),synset,subtaxonomies,!);(STR = 1,!)).

wordnet(DIR):-
  forall(background(_,WOR),loadprolog(WOR),loadcsv(PAL).

synset:-
  forall((ta(SIS,SNS,_,_),(cp(SIS,_,_);cp(_,SIS,_)),
  downcase_atom(SNS,LAD),useroutput(LAD),rootword(LAD,VTK,ROW,LEN)),
  (synset(SIS,ROW,LEN);synsetN(SIS,ROW,LEN);synsetF(SIS,VTK,ROW,LEN);
  synset1N(SIS,ROW,LEN);synset0N(SIS,LEN))).

subtaxonomies:-
  forall(cs(SIS,SIY),
  (findall(PAR,hypernym(SIY,PAR),TAT),list_to_set(TAT,CLEAN),
  append([SIY],CLEAN,SIYTAT),assert(ch(SIS,SIYTAT)))).
```

Listing 4: sp/0

Rule *wordnet*/1 is consulting the WordNet resources into Prolog (WOR) and the English dictionary in CSV file for removing irrelevant tokens and sequences (PAL). With rule *synset*/0 the related WordNet synset(s) is/are detected for each concept. To do so, the label is divided into single sequences, sequences not found in the dictionary are removed, and the length of the remaining root word is measured with rule *rootword*/4. This is necessary to define if the remaining root word consists of one or multiple sequences. Root words consisting of only one sequence are detected with *synset*1/3. Root words containing multiple sequences are searched in WordNet using *synsetN*/3. If no synset is found for the compound root word, each sequence is searched in WordNet with *synsetF*/4. Each matched synset is displayed along with its synonyms and its gloss to the user with *synsetU*/2. If no synset is found, *synset*1*N*/3, *synset*0*N*/2, or *synsetZ*/1 satisfies. If a single

sequence of the MWE should be ignored, or the complete label should be ignored, the user can skip the detection. For each verified synset, its (sub)-taxonomie are resulted using rule *subtaxonomies*/0, along with the included rule *hypernym*/2 for detecting the (inherited) hypernyms.

The matching to detect similarity between pairs is performed with *matching*/0, see Listing 5.

```
matching:-
  st(STR,_,BKL,BKG,BKP,UIV),
  ((STR = 1,matchlabels,((UIV=1,matchuser);UIV=0));(STR = 2,matchpath,((BKL=1,matchlabels);BKL=0),
  ((BKG=1,matchlanguage);BKG=0),((BKP=1,matchcontent);BKP=0),((UIV=1,matchuser);UIV=0))).

matchlabels:-
  forall(cp(SIS1,SIS2,SIP),(ta(SIS1,SNS1,SIP,_),ta(SIS2,SNS2,SIP,_),isub(SNS1,SNS2,true,ISS),
  assert(is(SIS1,SIS2,SIP,ISS)))).

matchlanguage:-
  forall(cp(SIS1,SIS2,SIP),(forall((cs(SIS1,SIY1),cs(SIS2,SIY2)),(tokenize(SIY1,NTK1),
  tokenize(SIY2,NTK2),jaccard(NTK1,NTK2,JAS),assert(ja(SIS1,SIS2,SIP,JAS)))))).

matchpath:-
  forall(cp(SIS1,SIS2,SIP),(forall((ch(SIS1,TAT1),ch(SIS2,TAT2))(member(ELE,TAT1),
  member(ELE,TAT2),index(IDX1,TAT1,ELE),index(IDX2,TAT2,ELE),PATH is (IDX1 + IDX2)-1,
  RAD is (1/PATH),assert(path(SIS1,SIS2,SIP,RAD)))))).

matchcontent:-
  forall(cp(SIS1,SIS2,SIP),(ta(SIS1,_,SIP,PRO1),ta(SIS2,_,SIP,PRO2),
  jaccard(PRO1,PRO2,PAS),assert(pr(SIS1,SIS2,SIP,PAS)))).

matchuser:-
  forall(cp(SIS1,SIS2,SIP),(ta(SIS1,SNS1,SIP,_),ta(SIS2,SNS2,SIP,_),
  (useroutput('Define if similar(0/1)'),useroutput(SNS1),useroutput(SNS2),read(UIV),
  assert(ui(SIS1,SIS2,SIP,UIV))))).
```

Listing 5: sm/0

The rule *matchlabels*/0 is performing a comparison based on the initial labels of the two concepts. Hereby, the built in predicate *isub*/4 is used. After the comparison, the predicate *is*/4 is created holding the comparison result ISS. The glosses are compared with using Jaccard similarity in the rule *matchlanguage*/0. To do so, each detected synset is matched with the WordNet gloss (GLO). For each gloss, stop words and punctmarks are removed before storing the comparison result in the fact *ja*/4. The (sub)-taxonomies of the synsets are compared in rule *matchpath*/0. Hereby, it first searches for the first common subsumer in both paths with utilizing the built-in predicate

member/2. Afterwards, it analyzes the position of the subsumer in both paths by exploiting the index with the built-in predicate *index*/3. Finally, both weights are substracted and the result is captured in the fact *se*/4. If the concepts are assigned with properties, a comparison between the two sets is performed using Jaccard similarity in rule *matchcontent*/0, before asserting the predicate *pr*/4 including the similarity result. With the rule *matchuser*/0, the user can, in addition to the involved techniques, also indicate if the concepts are similar. For each pair, the user can define if a matching exists (1), or not (0), which is later compared with the result(s) provided through the automatically performing matching techniques.

As each concept can have multiple synsets, and the strategy can also consist of different techniques, the combined results are detected with rule *combine*/0, see Listing 6.

```
combine:-
  st(STR,_,BKL,BKG,BKP,_),SUM is 1 + BKL + BKG + BKP,((STR = 1,combineall(STR,SUM));
  (STR = 2,combinestructure,((BKG=1,combinegloss);BKG=0),combineall(STR,SUM))).

combinestructure:-
  forall(cp(SIS1,SIS2,SIP),(findall(TAS,se(SIS1,SIS2,SIP,TAS),TASL),sumlist(TASL,SUM),
  length(TASL,LEN),SAC = SUM / LEN,assert(so(SIS1,SIS2,SIP,SAC)))).

combinegloss:-
  forall(cp(SIS1,SIS2,SIP),(findall(JAS,ja(SIS1,SIS2,SIP,JAS),JASL),sumlist(JASL,SUM),
  length(JASL,LEN),JAC = SUM / LEN,assert(jc(SIS1,SIS2,SIP,JAC)))).

combineall(STR,SUM):-
  forall(cp(SIS1,SIS2,SIP),((STR = 1, is(SIS1,SIS2,SIP,SIM),assert(re(SIS1,SIS2,SIP,SIM)));
  (STR = 2,so(SIS1,SIS2,SIP,SAC),((BKL=1,is(SIS1,SIS2,SIP,SIS));(BKL=0,SIS=0)),
  ((BKG=1,ja(SIS1,SIS2,SIP,JAC));(BKG=0,JAC=0)),((BKP=1,pr(SIS1,SIS2,SIP,JAK));
  (BKP=0,JAK=0)),SIM is (SIS + JAC + SAC + JAK)/SUM,assert(re(SIS1,SIS2,SIP,SIM))))).
```

Listing 6: sc/0

The results based on the structure of WordNet are combined with *combinestructure*/0. If the structure-based analysis is supported through analyzing the synset gloss, this results are combined in *combinegloss*/0. All results of different techniques are combined with *combineall*/1. Finally, the predicate *re*/4 is asserted, which includes the identifier of the super concept, the identifiers of the two compared sub concepts, and the combined similarity result SIM.

To filter if the pair is similar or not, the rule $sf/0$ is used, see Listing 7. To do so, the similarity result for each pair is compared against the in the strategy defined threshold. If user-involvement

is supported, the similarity result performed through the system and the user-rating is combined in $filteruser/2$. If user-involvement is not supported, only the similarity score detected by the system is compared in $filtermachine/2$. In both cases, the predicate $s/5$ is asserted.

```
filter:-
  st(_,THH,UIV),((sfu(THH,UIV),!);(sfn(THH,UIV),!)).

filteruser(THH,1):-
  forall((re(SIS1,SIS2,SIP,SIM),ui(SIS1,SIS2,SIP,UIV),SIU is ((SIM + UIV) / 2)),
  ((SIU >= THH,assert(s(SIP,SIS1,SIS2,THH,'yes'))));
  (SIU < THH,assert(s(SIP,SIS1,SIS2,THH,'no')))))).

filtermachine(THH,0):-
  forall(re(SIS1,SIS2,SIP,SIM),((SIM >= THH,assert(s(SIP,SIS1,SIS2,THH,'yes'))));
  (SIM< THH,assert(s(SIP,SIS1,SIS2,THH,'no'))))).
```

Listing 7: sf/0

Finally, each generated predicate $s/5$ detailing the similarity of the pair is output in CSV format using rule $output/0$, see Listing 8. If the assessment result should be used to extend the taxonomy, the rule $extend/0$ is used in addition, including two foregone rules. In rule $extendsiblings/0$, for each concept, the semantically similar sibling concepts are resulted using rule $siblingssimilar/2$. Afterwards, the user has to define a label generalising the included concepts, or the user can ignore the creation of the mediator concept, using rule $similarslabel/0$. Finally, each mediator concept is added to the knowledge base, respectively output to CSV file format using rule $extendoutput/0$.

```
output:-
  writeln('Define a path for output'),read(PAT),findall(s(SIP,SIS1,SIS2,THH,SA),
  (s(SIP,SIS1,SIS2,THH,SA)), Rows),write_csv(PAT, Rows).

extend:-
    extendsiblings,extendoutput.

extendsiblings:-
    ta(SIS,SNS,SIP,_),useroutput(SNS),siblingssimilar(SIS,NBS),similarlabel(SIS,NBS,MED),
    (not(MED=0),assert(m(SIS,NBS,SIP,MED)));(MED=0).

extendoutput:-
  useroutput('Define a path for mediators'),read(PAT),findall(m(SIS,NBS,MED),
  (m(SIS,NBS,MED)), Rows),writecsv(PAT, Rows).
```

Listing 8: so/0

4. Experimental Evaluation

To demonstrate the efficiency of *Taxo-Semantics* when comparing concepts in e-catalogs, the system is evaluated on three e-commerce databases, see Table 1. The Adventureworks, and the Northwind database are available through Microsofts hosting site CodePlex[5]. The *Festool* database is provided through a German retailing firm. All three databases are e-commerce applications, which are using taxonomies (i.e. e-catalogs) to categorize goods. For all three databases, the labels used to describe the concepts mainly consist of MWEs. The databases were used as provided with minimal modification. Concepts that have been deactivated by the marketing expert, i.e. not shown to the customers anymore, have been removed from the analysis, as given for the Festool database. In addition, labels in plural have been modified into its singular form, and for the matching with WordNet, all characters have been transferred to lowercase.

Table 1
Characteristics of the four databases used for experimental evaluation.

Characteristic	Adventureworks	Northwind	Festool
Number Root Concepts	1	1	1
Number Super Concepts	4	8	9
Number Sub Concepts	37	22	44
Average Mean of Sequences per Sub Concept	1.19 (±0.40)	1.32 (±0,48)	2.98 (±1.37)
Average Mean of Irrelevant Sequences per Sub	0 (±0)	0.18 (±0.39)	0,44 (±0.88)
Number of Properties assigned to Sub Concepts	48	-	132
Number of Concept Pairs	188	27	115
Number of Concept Pairs being similar	37	7	57
Number of Concept Pairs being dissimilar	151	20	58

Each taxonomy is considered to consist of three hierarchies (root concept, super concept, sub concept) and each sub concept contains maximum three properties. The properties for the Festool database are already provided. For the Adventureworks database, the properties were taken from the demo plattform Componentone[6]. No demo is provided for Northwind, thats why no properties are provided for this database. The task is to define the similarity between the concepts of the third hierarchy (sub-concept) analogues to the judgement provided by an expert user. The assessment is performed using WordNet. For example, the concepts inside the Adventureworks database "shorts" and "gloves" belong to the super concept "clothing". However, for both concepts there exists a more

[5]https://www.codeplex.com/
[6]http://www.componentone.com/

specific hypernym, namely "trousers" and "hand wear". Through this, "shorts" and "gloves" are considered as not similar. In contrary, a "bib-short" and "short" should be detected as similar as both have the common hypernym "trousers". The experiments are divided into three directions to provide the most comprehensive and reproducible analysis:

- TS-Label and TS-Bk (without user-involvement) are compared with other well-known related techniques, see Table 2. This helps to identify, which strategy performs best for which database characteristics. TS-Label is compared with other lexical-based approaches, namely the Hamming and Levenshtein distance. TS-Bk is compared with structure-based approaches, the Wu Palmer distance, the distance metrics presented by Ahsae, as well as the structure based methods presented by Li, and by Liu. Furthermore, the used structure-based technique inside TS-Bk is added to this category. To do so, TS-Bk is considered without the flexibility to add further techniques. According to the literature, it is in the follwing named Shortest Path Measure. Each technique is implemented in Prolog based on the formula presented in Lingling et al. (2013), and a cross-verification using WordNet Similarity for Java (WS4J[7]). Where necessary, the technique was normalized using the maximum length of the label/concept. Each technique was modified to be applicable for assessing MWEs, analogues to our strategies.

Table 2
Compared techniques used for assessing semantic similarity between concepts.

Technique/Strategy	Description
Levenshtein (1966)	lexical-based
Hamming	lexical-based
Shortes Path Measure	structure-based
Wu and Palmer (1994)	structure-based
Ahsae et al. (2012)	structure-based
Li et al. (2013)	structure-based
Liu et al. (2012)	structure-based
TS-Label	lexical-based
TS-BK	background-knowledge
TS-User	TS-BK, user-invovlement

- As TS-Bk allows to combine multiple techniques inside one strategy, different single variants

[7] http://ws4jdemo.appspot.com/

of the strategy are also investigated, see Table 3. This means that each combination of techniques is considered separately. This helps to identify, which combination of different techniques helps to improve the matching quality result. In addition to the techniques having the aim to semantically compare the sibling concepts, the provided translation service is also evaluated. To do so, the database that is provided not only in English (Festool) is used. More precisely, as Festool is a German company, we compare the translation of the German taxonomy into the English taxonomy performed by the expert, with the translation of the German taxonomy into the English taxonomy performed by the machine. To better distinguish between the results, we use the adequacy of the translation (all meaning, most meaning, much meaning, little meaning, none meaning).

Table 3
Characteristics of the subvariants used for strategy *Taxo-Semantics-BK*.

Variant	Label	Language	Structure	Content
TS-Bk-Label	yes	no	yes	no
TS-Bk-Content	no	no	yes	yes
TS-Bk-Language	no	yes	yes	no
TS-Bk-Textual	yes	yes	yes	no
TS-Bk-Complete	yes	yes	yes	yes

- The strategy using user-involvement, i.e. TS-User, is compared with TS-Bk. Through this, it can be stated, if involving the non-expert user can improve the matching quality result. Furthermore, as the best performing variant of TS-Bk is used for comparison, it states the minimal improvement to be expected. However, as the choice of the most related WordNet synset effects our strategies, we also investigate the probability to choose the correct synset. To do so, we firstly investigate the number of matched synsets shown to the expert, as well as shown to the non-expert. As the number of synsets shown to the users has an influence to choose the correct synset, but also has an influence on the time required to choose the correct synset, this measure states two characteristics. To better distinguish between the results, we compute the average number of synsets shown to the user per concept, but also investigate their distribution. Hereby, a concept resulting exactly one synset to be choosen is referred as 1:1 snyset, a concept resulting two synsets is referred as 1:2 synset, and so on. Secondly, we investigate the possible information loss when the expert has to assess the most related synset(s), compared to when the non-expert has to assess the most related synset(s). To do

so, we investigate the percentage of synsets being MWE, the percentage of synsets performing as compound synsets for a single label, as well as the percentage of single synsets for a single label. This is, because non-expert users tend to skip sequences, if they have already defined another synsets for another sequence.

- Finally, an analytical comparison, and a statistical significance measure is performed to highlight the strength and weakness of the presented decision support system. Hereby, the proposed strategies are compared with existing techniques in a theoretical manner by investigating the characteristics of the used techniques, benefits, and its drawbacks.

The three proposed matching strategies, its variants, as well as the above-mentioned related lexical-based and structure-based techniques, are evaluated with using the standard metrics used in information retrieval, namely the **Balanced Accuracy** (shorted as $BACC$), given in Equation 16:

$$BACC = \frac{TPR + TNR}{2}, \tag{16}$$

where TPR is the true positive rate, and TNR the true negative rate. The TPR, also referred as **Sensitivity**, measures the proportion of correctly identified positives, as given in Equation 17:

$$TPR = \frac{TP}{TP + FN}, \tag{17}$$

where TP is a true positive, and FN is a false negative statement. The TNR, also referred as **Specificity**, measures the proportion of correctly identified negatives, as given in Equation 18:

$$TNR = \frac{TN}{TN + FP}, \tag{18}$$

where TN is a true negative, and FP is a false positive statement. A statement is true, if the assessment provided through the system is equal to the assessment provided through the expert (positive, or negative). A statement is false, when both assessments are unequal. The maximum accuracy is 1, the minimum accuracy is 0. For each technique/strategy/variant, the average BACC is computed for each e-catalog to analyze the overall result, as given in Equation 19:

$$\overline{BACC} = \frac{BACC_{TH_1} + ... + BACC_{TH_N}}{N_{TH}}, \tag{19}$$

where $BACC$ is the balanced accuracy for a single threshold TH, and N is the number of thresholds. In total, 21 thresholds were investigated in steps of 0.05: 0.00, 0.05, 0.10, ..., 1.00. In addition, the $BACC$ for each of the 21 thresholds is computed and highlighted to analyze the consistency of the technique/strategy/variant (a graphical summary for the most meaningful thresholds is presented in Figure 9, 10, and 11). The improvement (shorted as INC) of the strategies compared to existing techniques is compared with analyzing the relative increase of balanced accuracy, as given in Equation 20:

$$INC = (\frac{\overline{BACC_{Strategy}}}{\overline{BACC_{Technique}}} * 100) - 100, \tag{20}$$

where $Strategy$ stands for the $\overline{BACC}$ result of a strategy of *Taxo-Semantics*, and $Technique$ stands for the $\overline{BACC}$ result of an existing technique. This formula is also used to measure the improvement when combining different techniques inside the strategy TS-Bk, as well as when using user-involvement in the strategy TS-User. For the latter, we follow the most recent progress of the OAEI for evaluating user-involvement. Namely, by defining a 10%, 20%, and 30% error rate for positive and negative statements. Hereby, the mentioned percentage of pairs is classified incorrectly by the users to simulate the by nature chance when involving non-experts. The average result of different users is finally taken as balanced accuracy.

4.1. Comparison of strategies with existing approaches

The obtained results evident, that for the three e-catalogs, the *Taxo-Semantics* matching strategies could outperform existing lexical- as well as structure-based techniques, see Figures 6, 7, and 8, and Table 4.

Table 4
Improvement Taxo-Semantics for different databases compared to existing techniques.

Existing	Taxo-Semantics	Adventureworks	Northwind	Festool	Average
Lexical-Based	TS-Label	1.14	11.04	9.73	7.31
Best	TS-Label	0.19	8.35	7.83	5.45
Structure-Based	TS-Bk	6.60	3.05	16.94	8.80
Best	TS-Bk	5,61	0.00	2.84	2.81
TS-User	TS-Bk	24.48	32.75	11.72	22.99

TS-Label could increase the accuracy compared to existing lexical-based techniques by on average +7.31 %. Using the ISub library instead of the best performing lexical-based technique Levenshtein results an increase of on average +5.45 % across all three databases. This improvement

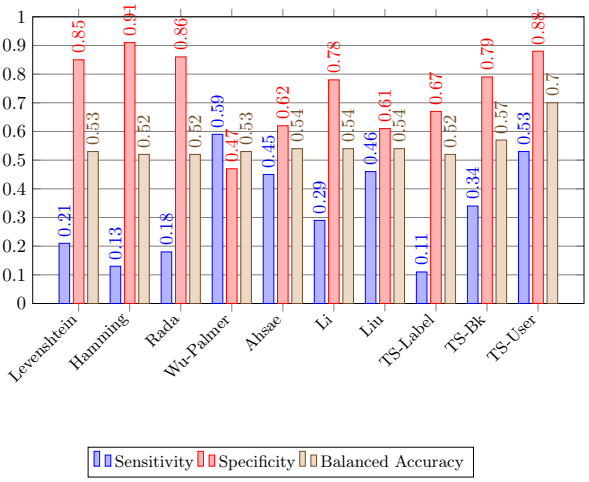


Fig. 6. Sensitivity, Specificity, and Balanced Accuracy results achieved for the Adventureworks e-catalog.

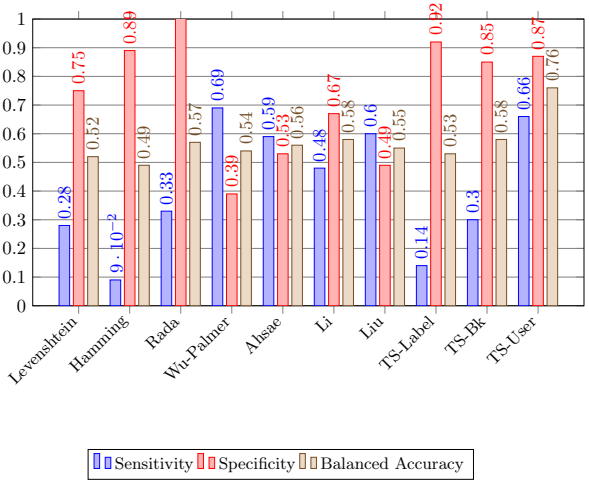


Fig. 7. Sensitivity, Specificity, and Balanced Accuracy results achieved for the Northwind e-catalog.

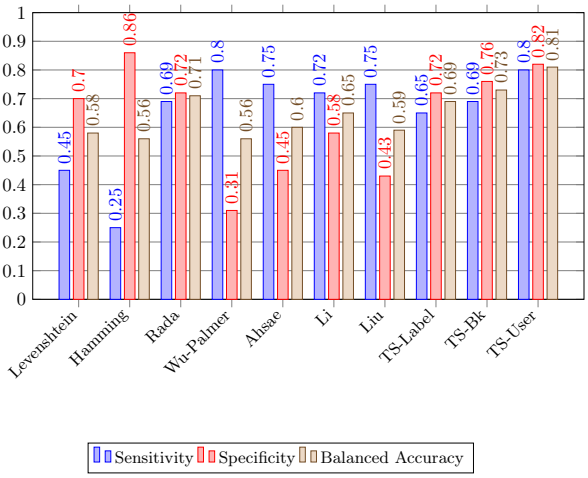


Fig. 8. Sensitivity, Specificity, and Balanced Accuracy results achieved for the Festool e-catalog.

Table 5
Improvement Taxo-Semantics when combining TS-Bk with additional techniques.

Technique	Subvariant	Adventureworks	Northwind	Festool	Average
Shortest Path	TS-Bk-Lexical	0.00	-4.39	0.00	-1.46
Shortest Path	TS-Bk-Content	8.65	-	-7.09	0.78
Shortest Path	TS-Bk-Language	-0.96	0.88	2.84	0.92
Shortest Path	TS-Bk-Textual	0.00	-1.75	1.42	-0.11
Shortest Path	TS-Bk-Complete	3.85	-	-2.13	0.86

is achieved as ISub is normalizing the two labels to be compared. More precisely, all sequences of MWEs are mapped to lowercase, and stopwords inside the labels are automatically removed. This is important, as in real-world databases, the labels often contain provider-specific abbreviations, or the labels only differ in having different adjectives before a noun. Consequently, the highest improvement was performed for the Norhtwind, and the Festool database. Both are using irrelevant sequences to label the concept, and Festool is in addition using provider-specific sequences.

TS-Bk could outperform existing structure-based techniques by on average +8.80 %. Compared to the best performing existing structure-based techniques, an increase of on average +2.81 % was performed. The highest improvement was achieved for the Festool database. This is because this database is using the highest portion of relevent sequences inside the MWEs. This results demonstrate the importance of semantically rich labels for sibling sub concepts. However, in contrast to the lexical-based techniques, not a single existing technique performed second-best. This shows

the inconsistency of existing techniques, which can be overcome by using TS-Bk, respectively a combination of multiple techniques. For the Adventurworks database, an increase of +5.61 % was performed compared to when only using the Shortest Path Measure. The obtained results for the Northwind database are equal to the Li distance result. And, for the Festool database, the accuracy could be increased by +2.84 % compared to the Shortest Path Measure. The results for the three databases in addition demonstrate that the best strcuture-based technique was used inside TS-Bk. When not combining the Shortest Path Measure with additional techniques, the an average balanced accuracy of 0.60 was achieved. The other structure-based techniques achieved a lower balanced accuracy: 0.54 (Wu and Palmer), 0.57 (Ahsae), 0.59 (Li), and 0.56 (Liu).

When comparing the strategy TS-Label with TS-Bk it is obvious that the lexical-based strategy shows almost the same balanced accuracy result. However, this is affected by a very high specificity compared to a low sensitivity. The reason for this is that the MWEs get more similar the deeper the concept occur inside the taxonomy. Through this, a lower result can be expected when using a higher level of the taxonomy for the experiments. In contrast, the structure-based techniques show a higher harmony when comparing the sensitivity and specificity.

4.2. Comparison of additional techniques inside combined strategies

When comparing the different variants of TS-Bk, multiple observations can be highlighted. Combining the lexical-based technique with structure-based techniques inside the variant TS-Bk-Label has not increased accuracy. The adding of a content-based analysis showed an inconsistent improvement. For the Adventureworks database, the accuracy could be increased by +8.65 %. For the Festool database, a decrease was performed. The reason is that the Festool database is using very similar properties across all concepts. Through this, a higher similarity result is computed, which negatively affects the accuracy. The Adventureworks database in contrast is using more different sets of properties to more precisely distinguish between concepts. The highest improvement was performed when combining the language-based technique with the strucuture-based technique inside TS-Bk-Language. On average a further decrease of on average +0.92 % was performed.

When summarizing the results regarding the translation service, it can be stated that most of the concepts are translated with all meaning or most meaning (26.19 %, and 35.71 %). One difficulty regarding the translation was that the Festool is using for some terms provider-specific translations, i.e. alternative synonyms. For that reason, some MWE have been translated with much meaning

(23.81 %). A small number of concepts has been translated with little or none meaning (9.52 %, and 4.76 %), which highlights that for most of the MWE a meaningful translation has been performed using Google translator.

4.3. Comparison of user-involvement with background-knowledge

The strategy *Taxo-Semantics-User* is based on the strategy *Taxo-Semantics-Bk*, but in addition supports user-involvement. The comparison between both strategies allows to state if user-involvement can additionally affect the matching quality result. To demonstrate the minimal expected increase of accuracy, the best performing subvariant of TS-Bk is compared with the accuracy result when taking the comparison result of non-experts into account. For all three databases, the adding of user-involvement could significantly improve the matching quality result.

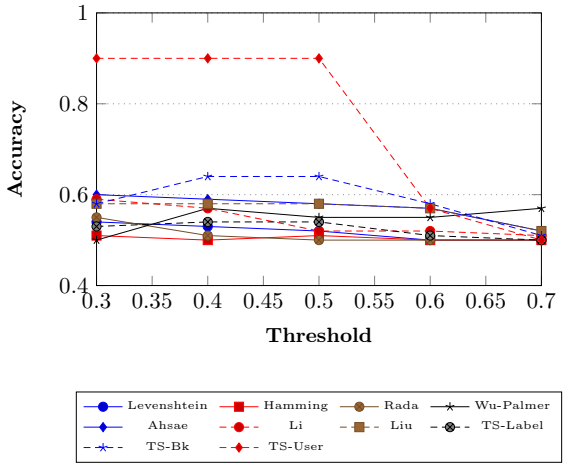


Fig. 9. Balanced Accuracy results for various thresholds achieved for the Adventureworks e-catalog.

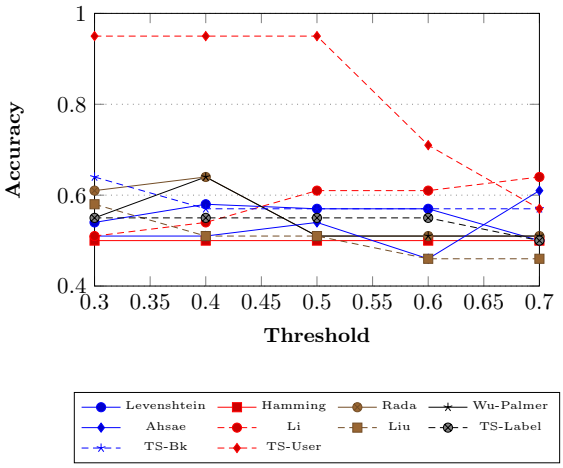


Fig. 10. Balanced Accuracy results for various thresholds achieved for the Northwind e-catalog.

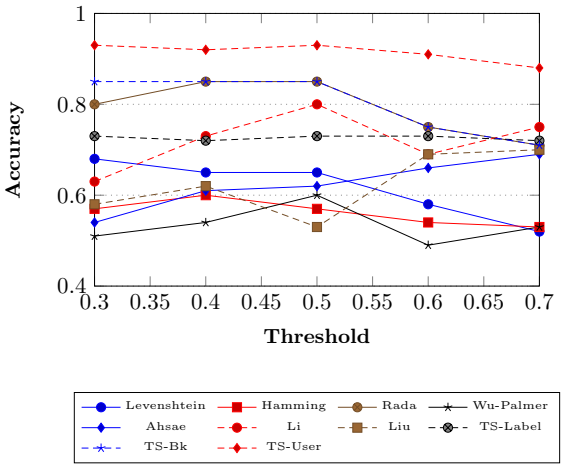


Fig. 11. Balanced Accuracy results for various thresholds achieved for the Festool e-catalog.

On average, an increase of +22.99 % was performed. The highest improvement was performed for the Northwind database (+32.75 %), as this database showed the lowest accuracy result across all existing techniques and proposed strategies. This highlights that supervised user-involvement not only increases the matching quality result itself, but also the consistency of the used strategy across different e-catalogs, see Figures 9, 10, and 11.

The probability to choose the correct synset goes from 40.29 % to 55.99 %, as shown in Table 6. For the Adventureworks and the Northwind database, maximum two sequences have been searched

for the concepts inside Wordnet. Depending on the position of the sequence, the probability further increases to choose the correct synset. For the Festool database, maximum three valid sequences have been searched for the concepts inside WordNet. Similar as for the other two databases, the probability increases with the position of the sequence. The possibility to increase the time required to choose the correct synset as the expert users, and to increase the probability would be to remember the choice for repeating sequences. For example in the Festool database, many concepts are using labels that overlap with other concepts and only differ in a single sequence. However, this would increase the cost of computation and decrease the flexibility of the decision support system.

When comparing the expert users with the non-expert users regarding the actually chosen synsets, it is further clear that the difference mainly consists in skipping sequences, see Table 7. As the non-expert users tend to identify the same MWE inside WordNet as the expert user, the MWE that can not be found in WordNet are slightly more error-prone. Through this, some compound terms are becoming labels consisting of a single sequence after the matching with WordNet.

Table 6
Probability for expert and non-expert users to choose the most related synset for three different databases.

Sequence/Synset		Adventureworks	Northwind	Festool
	Average	4.43 ±3.39	2.91 ±2.22	3.98 ±2.79
	1:1 - 1:3	56.75	63.63	52.39
	1:3 - 1:6	21.62	31.82	30.95
Synset 1	1:7 - 1:9	10.81	0.00	14.28
	1:10 - 1:12	8.11	4.55	2.38
	1:12 - 1:15	2.70	0.00	0.00
	Probability	40.29 ±31.81	55.99 ±35.75	42.77 ±32.49
	Average	7.00 ±5.59	1.33 ±0.58	2.94 ±2.09
	1:1 - 1:3	33.33	100.00	78.13
	1:3 - 1:6	16.67	0.00	15.63
Synset 2	1:7 - 1:9	0.00	0.00	3.13
	1:10 - 1:12	50.00	0.00	3.13
	1:12 - 1:15	0.00	0.00	0.00
	Probability	41.67 ±45.64	83.33 ±26.87	47.01 ±32.49
	Average	-	-	3.1 ±3.21
	1:1 - 1:3	-	-	80.00
	1:3 - 1:6	-	-	0.00
Synset 3	1:7 - 1:9	-	-	20.00
	1:10 - 1:12	-	-	0.00
	1:12 - 1:15	-	-	0.00
	Probability	-	-	63.89 ±39.65

Table 7
Information loss for non-expert users to choose the most related synset for three different databases.

User	Label	Adventureworks	Northwind	Festool
	MWE	2.70	5.00	0.00
Expert	Compound	13.51	0.00	45.45
	Single	83.78	95.00	54.54
	MWE	2.70	4.55	0.00
Non-Expert	Compound	0.00	0.00	36.36
	Single	97.30	95.45	63.63

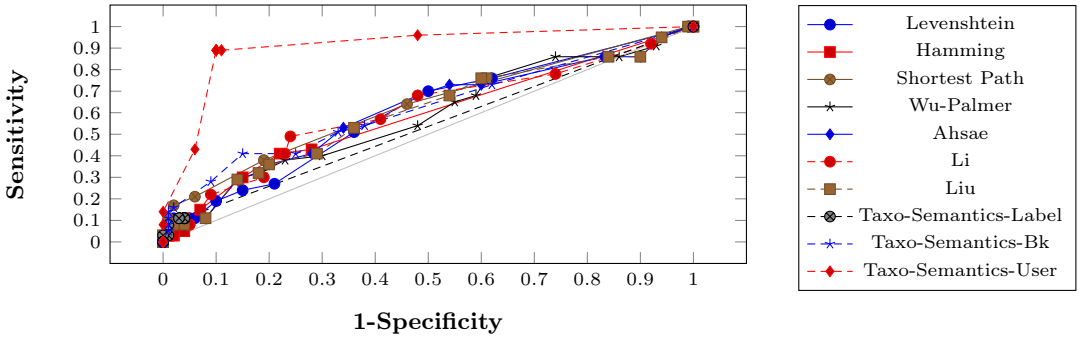
4.4. Analytical comparison and statistical significance

In addition to the metrics used in information retrieval, a statistical significance measure is performed using the Receiver Operating Characteristics (ROC) curve and the resulting Area Under Curve (AUC) measure. The ROC curve is comparing for each of the above mentioned thresholds, the false positive rate ($FPR = 1 - TNR$) against the TPR. Through this, the diagram can clearly identify if the technique/strategy performs well or badly in terms of a binary classification task. The AUC measure represents the area under the ROC curve. Consequently, a technique/strategy having a higher AUC measure is considered as better compared to a technique/strategy having a lower AUC area. To better distinguish between the different values, the traditional academic point system is used. It classifies an AUC between 0.9 and 1.0 as excellent, an AUC between 0.8 and 0.9 as good, an AUC between 0.7 and 0.8 as fair, an AUC between 0.6 and 0.7 as poor, and an AUC between 0.5 and 0.6 as fail.

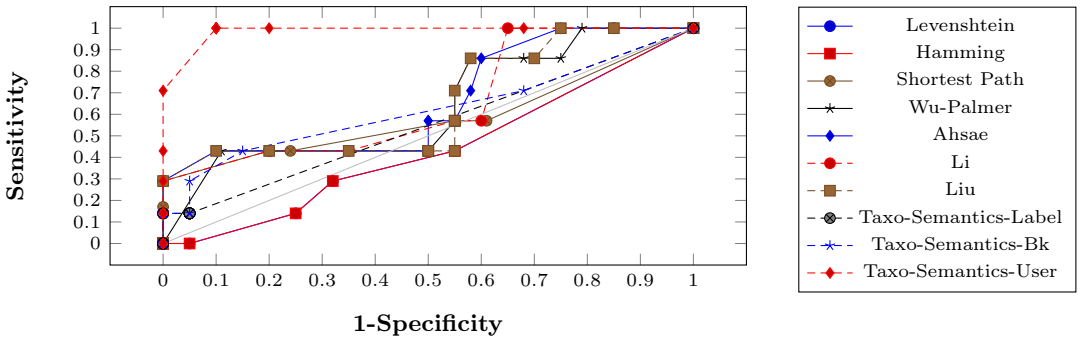
According to the statistical significance measure, the Shortest Path Measure, as used in TS-Bk, is the only single technique that shows a fair result, see Table 8. The lexical-based techniques and strategies show overall a poor result, or the technique fails. A better statistical result show the structure-based techniques and strategies. When comparing TS-Bk with the Shortest Path Measure, only a slight increase can be performed. The strategy using user-involvement is the only technique performing an excellent result.

Table 8
Semantic assessment area under curve comparison results for three databases.

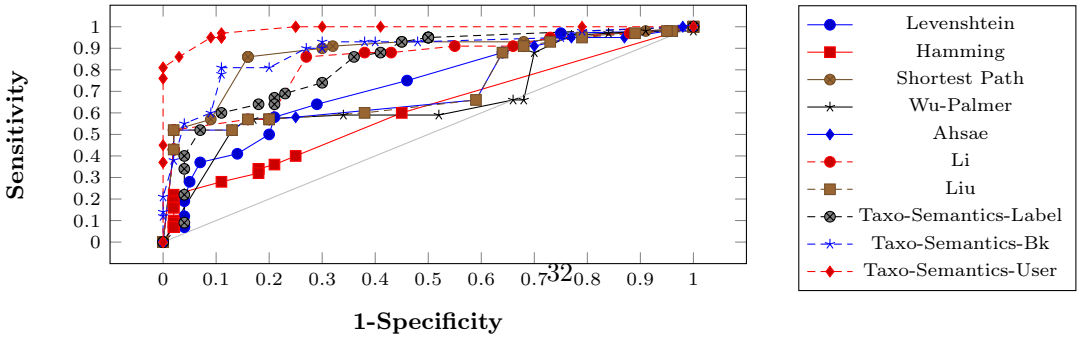
Technique/Strategy	Adventureworks	Northwind	Festool	Mean	Point
Levenshtein (1966)	0.61	0.51	0.73	0.62 ±0.11	poor
Hamming	0.59	0.43	0.61	0.54 ±0.10	fail
Shortes Path Measure	0.63	0.58	0.88	0.70 ±0.16	fair
Wu and Palmer (1994)	0.58	0.63	0.67	0.63 ±0.05	poor
Ahsae et al. (2012)	0.60	0.66	0.72	0.66 ±0.06	poor
Li et al. (2013)	0.61	0.65	0.83	0.70 ±0.12	fair
Liu et al. (2012)	0.60	0.65	0.72	0.66 ±0.06	poor
TS-Label	0.54	0.55	0.83	0.64 ±0.16	poor
TS-BK	0.62	0.62	0.89	0.71 ±0.16	fair
TS-User	0.90	0.99	0.99	0.96 ±0.05	excellent



(a) Adventureworks database



(b) Northwind database



(c) Festool database

Fig. 12. Semantic assessment receiver operating characteristic comparison results for three databases.

From an analytical perspective, the Hamming and Levenshtein distance are analyzing the similarity based on the strings of the labels, as similar to the strategy proposed in *Taxo-Semantics-Label* Levenshtein (1966). The Hamming distance describes the number of characters being different at the same index. Each substitution necessary to transform the initial character into the target character increases the distance result (Li (2009)). The Levenshtein distance characterizes the minimum number of edits required to transform one string into the other. Each edit is qualified by the kind of transformation necessary: if a single character has to be added (insertions), deleted (deletions), or substituted (substitutions). Compared to the technique in *Taxo-Semantics-Label*, the Hamming distance requires that the strings to be compared have the same length. Consequently, if the strings do not have the same length, the strings have to be normalized, which results further computational costs, and a decrease of the similarity value. However, if the strings merely differ in adjectives or provider-specific addition sequences (e.g. "Car" compared to "Car Big"), the computation is meaningful, especially when weighting the adjectives lower than the main word. When comparing the Levenshtein distance with *Taxo-Semantics-Label* the difference is smaller as the used Isub[8] library is based on the Levenshtein distance Levenshtein (1966). However, the Isub library is capable to automatically convert each character into lower case. Through this, characters are considered as similar, because they have the same semantic weight. The distance measures presented in Wu and Palmer (1994); Ahsae et al. (2012); Li et al. (2013); Liu et al. (2012) are structure-based techniques. The Wu and Palmer (1994) measure is referred as edge-based measure, because of it is using the latest common subsumer of the two concepts to be compared, respectively the relative depth of the concepts. This is important for taxonomies consisting of multiple levels, further arbitrary relationships (e.g. meronyms), and if each concept belongs to exact one super concept. However, e-commerce are usually limited to the is-a knowledge. In addition, e-catalogs mainly consist of maximum four levels with redundant sub concepts. The technique proposed in Ahsae et al. (2012) is built upon the before-mentioned measure, but in also measures the difference of the depth. In addition, their attempt is capable to weight each sequence of a MWE different. However, in e-commerce this would require that the user has a high expertise about the domain, or further user-involvement would be required, which effects further computational costs. The Li et al. (2013) is mapping the concepts into a concept space. This has the benefit that the context of the term can be considered over a

[8] http://www.swi-prolog.org/

larger semantic network. However, the attempt does not consider to use the knowledge also for other measuers (e.g. language-based), as provided in *Taxo-Semantics-Bk*. The work presented in Liu et al. (2012) combines the shortest path measure with the depth of the concepts in a non-linear function, and is the most related technique to the work presented at hand. Their attempt is distinguishing between background knowledge consisting of merely the is-a structure (tree), and between background knowledge including more arbitrary relationships (graphs). However, as background knowledge, merely WordNet is used.

5. Conclusions

This work presented *Taxo-Semantics*, a system to assess similarity between concepts in e-catalogs, and to use the assessment result for extending e-catalogs. *Taxo-Semantics* differs from existing approaches in five directions. Firstly, MWEs can be analyzed through providing three different matching strategies. Secondly, the system includes a translation service for matching concepts not expressed in English for being matched to WordNet. Thirdly, the proposed system can input/-export the taxonomies, respectively the assessment result in relational database format. Fourthly, *Taxo-Semantics* provides optional user-involvement to support the matching with WordNet, and to affect the assessment result. Fifthly, the assessment result can be used to create mediator concepts for semantically similar sibling concepts. The extensive computational experiments performed on three e-catalogs, highlight the efficiency of *Taxo-Semantics* and the supported strategies. On average, the accuracy could be increased by +7.31 %, +8.80 %, and +22.99 % for the lexical-based, background-based, and the strategy supporting user-involvement.

Future work on *Taxo-Semantics* can be divided in three directions. Firstly, standard and upper-level taxonomies like eCl@ss or GS1 could be used. Those provide a more specific view on concepts across e-commerce domains and provide a further description for the concepts. This would allow to enrich the matching process by integrating a further language-based analysis in addition to using the WordNet gloss. Secondly, as the single sequences of MWEs usually have a different semantically weight, this can be considered for the background-based strategy. And thirdly, as taxonomies suffer from semiotic heterogeneity, i.e. misinterpretation of concepts, implicit feedback could be used to further analyze the goodness of the semantics of the taxonomy.

Acknowledgement

Thanks to Festool GmbH & Co. KG for providing real-world data to analyse in our experiments.

References

Ahsae, M.G., Naghibzadeh, M., Yasrebi Naeini, S.E., 2012. Semantic similarity assessment of words using weighted wordnet. International Journal of Machine Learning and Cybernetics 5, 479–490.

Aiken, M., Park, M., Simmons, L., Lindblom, T., 2009. Automatic translation in multilingual electronic meetings. Translation Journal 13.

Angermann, H., Ramzan, N., 2016a. E-commerce with smartstore.net - a review of the through microsoft distributed shop-system. Visual Studio One 44, 39–42.

Angermann, H., Ramzan, N., 2016b. Preference analysis with microsoft social engagement - big data and web 3.0 inside microsoft dynamics crm. Visual Studio One 44, 39–42.

Angermann, H., Ramzan, N., 2016c. Taxo-publish: Towards a solution to automatically personalize taxonomies in e-catalogs. Expert Systems with Applications 66, 76–94.

Cheatham, M., Dragisic, Z., Euzenat, J., Faria, D., Ferrara, A., Flouris, G., others, 2016. Results of the ontology alignment evaluation initiative 2015, in: Proceedings of the 10th International Workshop on Ontology Matching, CEUR Workshop Proceedings, Aachen, Germany. pp. 60–115.

Chuang, S., Chien, L., 2003. Enriching web taxonomies through subject categorization of query terms from search engine logs. Decision Support Systems 35, 113–127.

Faria, D., Martins, C., Nanavaty, A., Taheri, A., Pesquita, C., Santos, E., others, 2014. Agreementmakerlight results for oaei 2014, in: Proceedings of the 9th International Workshop on Ontology Matching, CEUR Workshop Proceedings, Aachen, Germany. pp. 1–8.

Fellbaum, C., 1998. WordNet: An Electronic Lexical Database. MIT Press, Cambridge, CB, USA.

Furlan, B., Batanovic, V., Nikolic, B., 2013. Semantic similarity of short texts in languages with a deficient natural language processing support. Decision Support Systems 55, 710–719.

Gomez-Perez, A., Fernández-López, M., Corcho, O., 2006. Ontological Engineering: with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web. Springer Science & Business Media.

Hogenboom, A., Heerschop, B., Frasincar, F., Kaymak, U., de Jong, F., 2013. Multi-lingual support for lexicon-based sentiment analysis guided by semantics. Decision Support Systems 62, 43–53.

Jimenez-Ruiz, E., C. Grau, B., Xia, W., Solimando, A., Chen, X., Cross, V., others, 2014. Logmap family results for oaei 2014, in: Proceedings of the 9th International Workshop on Ontology Matching, CEUR Workshop Proceedings, Aachen, Germany. pp. 1–9.

Kim, D., Wang, H., Oh, A.H., 2013. Context-dependent conceptualization, in: Proceedings of the 23rd international joint conference on Artificial Intelligence, Association for the Advancement of Artificial Intelligence Press, Palo Alto, CA, USA. pp. 2654–2661.

Levenshtein, W., 1966. Binary codes capable of correcting deletions, insertions and reversals. Soviet Physics Doklady (Springer) 10, 707–710.

Li, P., Wang, H., Zhu, K.Q., Wang, Z., Wu, X., 2013. Computing term similarity by large probabilistic isa knowledge, in: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, Association for Computing Machinery, New York City, NY, USA. pp. 1401–1410.

Li, S.Z., 2009. Encyclopedia of Biometrics. Springer Publishing Company, Incorporated, Heidelberg/Berlin, Germany.

Lingling, M., Runqing, H., Junzhong, G., 2013. A review of semantic similarity measures in wordnet. International Journal of Hybrid Information Technology 6, 1–12.

Liu, H., Bao, H., Xu, D., 2012. Concept vector for semantic similarity and relatedness based on wordnet structure. Journal of Systems and software 85, 370–381.

Ma, Y., Liu, J., Yu, Z., 2013. Concept name similarity calculation based on wordnet and ontology. Journal of Software 8, 746–753.

Meijer, K., Frasincar, F., Hogenboom, F., 2014. A semantic approach for extracting domain taxonomies from text. Decision Support Systems 62, 78–93.

Miller, G.A., 1995. Wordnet: A lexical database for english. Communications of the ACM 38, 39–41.

Nguyen, T., Conrad, S., 2013. A semantic similarity measure between nouns based on the structure of wordnet, in: Proceedings of International Conference on Information Integration and Web-based Applications & Services, Association for Computing Machinery, New York City, NY, USA. pp. 605–609.

Peukert, E., Eberius, J., Rahm, E., 2012. A self-configuring schema matching system, in: Proceedings of the 28th IEEE International Conference on Data Engineering, Institute of Electrical and Electronics Engineers, New York City, NY, USA. pp. 306–317.

Rada, R., Mili, H., Bicknell, E., Blettner, M., 1989. Development and application of a metric on semantic nets. Transactions on Systems, Man, and Cybernetics 19, 17–30.

Sag, I.A., Baldwin, T., Bond, F., Copestake, A., Flickinger, D., 2002. Multiword expressions: A pain in the neck for nlp, in: Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing, Springer Publishing Company, Incorporated, Berlin/Heidelberg, Germany. pp. 1–15.

Sanchez, D., Batet, M., 2013. A semantic similarity method based on information content exploiting multiple ontologies. Expert Systems with Applications 40, 1393–1399.

Sharma, A., Dey, S., 2015. Mining marketing intelligence from online reviews using sentiment analysis. International Journal of Intercultural Information Management 5, 57–82.

Shvaiko, P., Euzenat, J., 2013. Ontology matching: State of the art and future challenges. Transactions on Knowledge and Data Engineering 25, 158–176.

Wu, Z., Palmer, M., 1994. Verbs semantics and lexical selection, in: Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics, Association for Computing Machinery, New York City, NY, USA. pp. 133–138.