A Topic-centric Approach to Detecting New Evidences for Evidence-based Medical Guidelines

Qing Hu[1,2], Zhisheng Huang[1], Annette ten Teije[1], Frank van Harmelen[1], M. Scott Marshall[3] and Andre Dekker[3]

[1]*Department of Computer Science, VU University Amsterdam, De Boelelaan 1081, Amsterdam, The Netherlands*

[2]*College of Computer Science and Technology, Wuhan Univesity of Science and Technology, Wuhan, China*

[3]*Department of Radiation Oncology (MAASTRO), Maastricht University Medical Centre, Maastricht, The Netherlands*

Keywords: Evidence-based Medical Guidelines, Medical Guideline Update, Semantic Distance, Context-awareness, Topic-centric Approach.

Abstract: Evidence-based Medical guidelines are developed based on the best available evidence in biomedical science and clinical practice. Such evidence-based medical guidelines should be regularly updated, so that they can optimally serve medical practice by using the latest evidence from medical research. The usual approach to detect such new evidence is to use a set of terms from a guideline recommendation and to create queries for a biomedical search engine such as PubMed, with a ranking over a selected subset of terms to search for relevant new evidence. However, the terms that appear in a guideline recommendation do not always cover all of the information we need for the search, because the contextual information (e.g. time and location, user profile, topics) is usually missing in a guideline recommendation. Enhancing the search terms with contextual information would improve the quality of the search results. *In this paper, we propose a topic-centric approach to detect new evidence for updating evidence-based medical guidelines as a context-aware method to improve the search*. Our experiments show that this topic centric approach can find the goal evidence for 12 guideline statements out of 16 in our test set, compared with only 5 guideline statements that were found by using a non-topic centric approach.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

1 INTRODUCTION

Medical guidelines, or alternatively clinical guidelines, are conclusions or recommendations on the appropriate treatment and care of people with specific diseases and conditions, which are designed by medical authorities and organizations. Evidence-based medical guidelines are developed based on the best available evidence in biomedical science and clinical practice. Guideline recommendations in evidence-based medical guidelines are annotated with their underlying evidence and their evidence classes. Medical guidelines have been proved to be valuable for clinicians, nurses, and other healthcare professionals (Woolf et al., 1999)[1].

Ideally, a guideline should be updated immediately after new relevant evidence is published, so that the updated guideline can serve medical practice using the latest medical research. However, because of the sheer volumes of medical publications, the update of a guideline is often lagging behind medical scientific publications. Not only are the number of medical articles and the size of medical information very large, but also they are updated very frequently. For example, PubMed[2] alone contains more than 24 million citations for biomedical literature from MEDLINE[3]. Thus, updating a guideline is laborious and time-consuming.

In order to solve those disadvantages, some approaches have been proposed that use information retrieval or machine learning technology to find relevant new evidence automatically. Reinders et al. (Reinders et al., 2015) described a system to find relevant new evidence for guideline updates. The approach is based on MeSH terms and their TF-IDF weights, which results in the following disadvantages: i)the use of MeSH terms terms means that if a guideline

[1]https://en.wikipedia.org/wiki/Medical_guideline

[2]http://www.ncbi.nlm.nih.gov/pubmed

[3]http://www.nlm.nih.gov/bsd /pmresources.html

Hu, Q., Huang, Z., Teije, A., Harmelen, F., Marshall, M. and Dekker, A.
A Topic-centric Approach to Detecting New Evidences for Evidence-based Medical Guidelines.
DOI: 10.5220/0005698902820289
In *Proceedings of the 9th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2016) - Volume 5: HEALTHINF*, pages 282-289
ISBN: 978-989-758-170-0
Copyright © 2016 by SCITEPRESS – Science and Technology Publications, Lda. All rights reserved

statement does not use any MeSH term, there is no way to measure the relevance of a publication, ii)the use of TF-IDF weights means that the system has to gather all relevant sources, which is time-consuming, iii)the number of returned relevant articles is sometimes too large (sometimes even a few million), so that it is impossible for an expert to check if an evidence is really useful for the guideline update. Iruetaguena et al. (Iruetaguena et al, 2013) also developed an approach to find new evidence. That method is also based on gathering all relevant articles by searching the PubMed website, and then uses the Rosenfeld-Shiffman filtering algorithm to select the relevant articles. The experiment of that approach shows the recall is excellent, but the precision is very low (10.000 articles contain only 7 goal articles) (Reinders et al., 2015). In (Hu et al., 2015), we propose a method that uses a semantic distance measure to automatically find relevant new evidence for guideline updates. The advantage of using semantic distance is that the relevance measure can be achieved via the co-occurrence of terms in a biomedical article, which can be easily obtained via a biomedical search engine such as PubMed, instead of gathering a large corpus for the analysis.

The existing approaches to detect relevant evidence for guideline updates are using the terms appearing in a guideline statement. However, these terms appearing in a guideline statement do not always cover all of the information we need for the search, because the contextual information (e.g. time and location, user profile, topics) is usually missing in a guideline statement. Enhancing the relevance checking with contextual information would improve the quality of the search results. *In this paper, we propose a topic-centric approach to detect new evidence for updating evidence-based medical guidelines as a context-aware method to improve the search.* We consider the title of the section or subsection containing a guideline statement as the topic of that guideline statement. In the semantic distance based approach, the terms appearing in the topic (i.e., in the title of the section or subsection) should be ranked as more important than other terms. We have conducted several experiments with this topic-centric approach to find new relevant evidence for guideline updates. We will show that this topic-centric approach indeed provides a better result.

This paper is organized as follows: Section 2 introduces the basic structure of guidelines and the procedure of guideline update, presents the approach based on a semantic distance measure over terms, and describes several strategies using the semantic distance measure for finding new and relevant evidence for guidelines. Section 3 proposes the topic centric approach. Section 4 presents several experiments of our method on the update of guidelines. Section 5 discusses future work and concludes.

2 EVIDENCE-BASED GUIDELINES AND GUIDELINE UPDATES

2.1 Guideline Updates

Evidence-based medical guidelines are based on published scientific research findings. Those findings are usually found in medical publications such as those in PubMed. Selected articles are evaluated by an expert for their research quality, and are graded for the degree to which they contribute evidence using a classification system (NSRS, 2006).

A usual classification of research results in evidence levels consists of the following five classes (NSRS, 2006; NABON, 2012): Type A1: Systematic reviews, or that comprise at least several A2 quality trials whose results are consistent; Type A2: High-quality randomised comparative clinical trials of sufficient size and consistency; Type B: Randomised clinical trials of moderate quality or insufficient size, or other comparative trials (non-randomised, comparative cohort study, patient control study); Type C: Non-comparative trials, and Type D: Opinions of experts. Based on this classification of evidence, we can classify the conclusions in a guideline (sometimes called *guideline items*) with an evidence level. The following evidence levels for guideline items are proposed in (NABON, 2012): Level 1: Based on 1 systematic review (type A1) or at least 2 independent A2 reviews; Level 2: Based on at least 2 independent type B reviews; Level 3: Based on 1 type A2 or B research, or any level of C research, and Level 4: Opinions of experts.

Here is an example of a conclusion in a guideline in (NABON, 2012):

```
Classification: Level 1
Statement:
    The diagnostic reliability of ultrasound
    with an uncomplicated cyst is very high.
Evidence: A1 Kerlikowske 2003, B Boerner 1999,
    Thurfjell 2002, Vargas 2004
```

which consists of a conclusion classification ('Level 1'), a guideline statement, and its evidence items with one item classified as A1 and three items classified as B (jointly justifying the Level 1 of this conclusion).

In order to check if there is any new evidence from a scientific paper which is relevant to the guideline

statement, a natural way to proceed is to use the terms which appear in the guideline statement to create a query to search over a biomedical search engine such as PubMed. In our experiments reported in (Hu et al., 2015), we use Xerox's NLP tool (Ait-Mokhtar et al., 2013; Aït-Mokhtar et al., 2002) to identify the medical terms from UMLS and SNOMED CT which appear in guideline statements (Huang et al., 2014), and then use these terms to construct a PubMed query to search for relevant evidence. The resulting PubMed ID (alternatively called PMID) can serve as the ID of a retrieved evidence. A naive approach to creating such a PubMed query is to construct the conjunction or disjunction of all terms that appear in a guideline item. We have observed the following facts: i) the result size of the conjunctive query often leads to 0 results (67% of the cases), and ii) the result of the disjunctive query would frequently lead to too many results (average 812,632, max 9,211,547) (Reinders et al., 2015). The main problem of those approaches is that the semantic relevance of the terms is not well considered. An improved approach is to use a semantic distance measure to create search queries in which more relevant terms are preferred to less relevant terms. In other words, the semantic distance measure provides us with a method to rank the terms in the search query.

The method consists of several steps that need to be executed in order, as follows:

1. Extract the terms and the PMID of the evidences.
2. Use different terms ranking strategies.
3. Construct a PubMed query based on ranked terms.
4. Execute the query and evaluate the results.
5. Present the best results to the user.

In (Hu et al., 2015), we propose a semantic distance measure to rank terms for finding relevant evidence from a Biomedical search engine such as PubMed. Our semantic distance measure is based on the (widely shared) assumption that more frequently co-occurring terms are more semantically related.

In order to make this paper self-contained, we describe the relevant notions of the semantic distance method in the following:

The equation for our Normalized PubMed Distance (NPD) is as follows:

$$NPD(x,y) = \frac{max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log M - min\{\log f(x), \log f(y)\}}$$

Where $f(x)$ is the number of PubMed hits for the search term x; $f(y)$ is the number of PubMed hits for the search term y; $f(x,y)$ is the number of PubMed hits for the search terms x and y; M is the number of PMIDs indexed in PubMed (where M=23,000,000 at the time of writing). $NPD(x,y)$ can be understood intuitively as the symmetric conditional probability of co-occurrence of the search terms x and y (Cilibrasi and M.B.Vitanyi, 2007).

Let G be a set of guideline statements and $Terms$ be the set of all terms. The function $T : G \rightarrow Powerset(Terms)$ assigns a set of terms to each guideline statement such that $T(g)$ is the set of terms which appear in the guideline statement g. For each guideline statement $g \in G$ and a term x in $T(g)$, we define $AD(x,g)$ as the average distance of x to other terms in g:

$$AD(x,g) = \frac{\sum_{y \in T(g), y \neq x} NPD(x,y)}{|T(g)| - 1}$$

We define the *center term* $CT(g)$ as the term whose average distance to other terms (in the guideline statement g) is minimal:

$$CT(g) = arg_x \; min(AD(x,g))$$

We can now consider the following different strategies for term ranking:

- *Average Distance Ranking*(ADR): ranks the terms by their average distance value.
- *Central Distance Ranking*(CDR): ranks the terms by their distance to the center term, where the central distance of a term x in a guideline statement g, written as $CD(x,g)$, is defined as:

$$CD(x,g) = NPD(x, CT(g))$$

We propose the following criteria the for evaluting the results:

- Term Coverage Criteria: The more terms which appear in the guideline statement are used for search, the more relevant the results are;
- Evidence Coverage Criteria: The more original evidences have been covered in the search, the more relevant the results are;
- Bounded Number Criteria: It is not meaningful to have too many results (for example more than 10,000 papers). Furthermore, we would likely miss many evidence items if there are too few results (for example, less than 10 papers). Thus, we can set the upper bound and lower bound of the results. The former is called the upper bound number P_u, whereas the latter is called the lower bound number P_l.

Based on the three assumptions above, we design a heuristic function $f(i)$ to evaluate the search results at each step at the workflow above:

$$f(i) = k_1 T(i)/T + k_2 E(i)/E + k_3 (P_u - P(i))/P_u$$

where T is the total number of terms in the guideline statement; $T(i)$ is the number of selected terms in this search i; E is the total number of the evidence items for the guideline statement; $E(i)$ is the number of the original evidence items which has been covered in this search i; P_u is the upper bound number; $P(i)$ is the number of PMID's that result from this search i, if $P(i)$ is a number between P_u and P_l, and k_1, k_2, k_3 are the weights of the different criteria. It is easy to see that the first part of the heuristic function (e.g.,$k_1T(i)/T$) measures the Term Coverage Criterion, the second part of the function (e.g., $k_2E(i)/E$) measures the Evidence Coverage Criterion, whereas the third part of the function (e.g., $k_3(P_u - P(i))/P_u$) measures the Bounded Number Criterion with the meaning that the fewer results are returned, the more preferred they are (if the result size is between P_u and P_l).

In (Hu et al., 2015), we have reported several experiments to evaluate the above approach. We selected the Dutch breast cancer guideline (version 1.0, 2004) (NABON, 2004) and the Dutch breast cancer guideline (version 2.0, 2012) (NABON, 2012) as the test data. From these experiments, we found that there is room to improve the search results by reducing the sizes of the returned results and to find more goal evidence for more guideline items. In (Hu et al., 2015) and as explained above, the center term is defined as the term for which the average distance to other terms is minimal. That definition of center term is independent of the topic of a selected guideline conclusion (where by topic, we mean the titles of the sections or subsections in which the guideline conclusions are contained). An intuitive approach is to select the terms which appear in the topic to be a center term. The contribution of this paper is to develop this topic-centric approach to find new evidence. We will report the experiments that compare the non-topic-centric approach with the topic-centric approach in Section 4.

3 TOPIC-CENTRIC APPROACH FOR FINDING NEW EVIDENCES

Contextualization has been considered to be a useful approach to improve the quality of search, because the context can provide more precise information for users to make queries and to reduce the size of search results (Stalnaker, 1999). Typically, spatial and temporal information about the users and the systems are considered as contextual information, because they are usually not stated explicitly when users make a search. Personalization can be also considered as a special case of contextualization. The same scenario can be also applied to the topic that the search is concerned with, since this is usually also not stated explicitly.

For medical guidelines, it is quite convenient to obtain this topic information, because each guideline recommendation or conclusion is always covered in a section or a subsection with a specific title. Of course, the title of a section or a subsection may contain multiple terms. Again we can use the semantic distance measure to rank the terms appearing in the topic. Therefore, a topic centric approach to rank the terms can be done as follows:

1. Obtain the terms which appear in the title of section or subsection of a guideline conclusion. They are called the topic terms.
2. Rank the topic terms by using the semantic distance measure. The first term in the ranking is considered to be the center term.
3. Add non-topic terms which appear in the guideline statement one by one, based on their semantic distance to the center term.
4. Create a search query based on the merged set of the topic terms and non-topic terms.
5. Search over PubMed to find relevant evidence by using the generated queries.
6. Select the best query answer based on the heuristic function.

Let G be a set of guideline statements and $Terms$ be the set of all terms. The function $Topic : G \rightarrow Powerset(Terms)$ assigns a set of terms to each guideline statement such that $Topic(g)$ is the set of terms which appears in the title of the section or the subsection in which the guideline statement g appears. Of course, the intersection of the terms and the topic terms of a guideline statement may not be an empty set:

$$T(g) \cap Topic(g) \neq \emptyset$$

In the topic-centric approach, for each guideline statement $g \in G$ and term x in $Topic(g)$, we can define the average distance of term $x \in Topic(g)$, written as $AD_T(x,g)$ as follows:

$$AD_T(x,g) = \frac{\sum_{y \in Topic(g), y \neq x} NPD(x,y)}{|Topic(g)| - 1}$$

We define the *center term* $CT_T(g)$ in the topic as the term whose average distance to other terms (in the guideline statement g) is minimal:

$$CT_T(g) = arg_x \, min(AD_T(x,g))$$

In this paper, we use the strategy of the Central Distance Ranking with the topic. Namely, this strategy ranks the topic terms by their distance to the center term in the topic first, then ranks those non-topic terms by their distance to the center term in the topic, where the central distance of a term $x \in T(g)$ in the topic for a guideline statement g, written as $CD_T(x,g)$, is defined as:

$$CD_T(x,g) = NPD(x, CT_T(g))$$

4 EXPERIMENTS AND EVALUATION

We have implemented the guideline update tool as a component in SemanticCT, a semantically-enabled system for clinical trials (Huang et al., 2013; Hu et al., 2014)[4]. We have conducted several experiments for finding relevant evidence for guideline updates. We selected the Dutch breast cancer guideline (version 1.0, 2004) (NABON, 2004) and the Dutch breast cancer guideline (version 2.0, 2012) (NABON, 2012) as test data. For our experiments we have selected 16 conclusions which appear in both versions of the guidelines. Thus, the evidence items appearing in the second version of the guideline can serve as a gold standard to test the proposed approach in this paper. Namely, we want to know whether or not finding relevant evidence for the first version of the guideline can really find the target evidence (alternatively called *goal evidence items*) which was used on the second version of the guideline. For the non-topic-centric approach, one of our experiments in (Hu et al., 2015) is using the Central Distance Ranking. We compare the results of the topic centric approach with the results of the non-topic centric approach to see whether or not it can get a better result, namely, finding goal evidence items for more guideline statements.

Our first experiment is to use the topic centric approach to find relevant evidence for the sixteen selected guideline conclusions and use the same heuristic function to guide the search with the same weights on the three criteria, namely $k_1 = k_2 = k_3 = 1/3$ and the upper bound number $P_u = 1000$ and the lower bound number $P_l = 25$. The results of a comparison the non-topic-centric approach are shown in Table 1. In this experiment, the topic centric approach can find the goal evidence items for 12 guideline statements out of 16 ones. Compared with the results obtained by using the non-topic centric approach (which can find goal evidences for only 5 guideline statements), it gets

[4]http://wasp.cs.vu.nl/sct

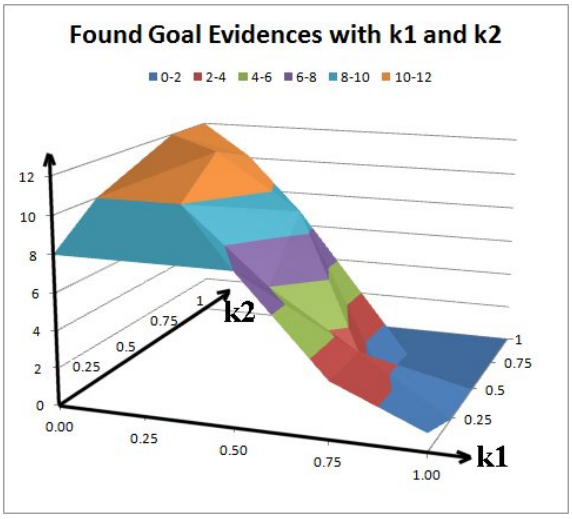

Figure 1: Systematic Tests on Different Weights.

a much better result. We have also observed that it is not always the case that the topic-centric approach gives a better result. For example, for the guideline statement 04_1_3, the non-topic-centric approach can find 3 goal evidence items, whereas the topic-centric approach can find only one. This increase in items for which any goal evidence is found (12 against 5) comes at the price of a slightly lower percentage of evidence items found per case (dropping from 57% to 46%). This small drop is well worth the steep increase from 5 to 12 cases for which any goal evidence is found. As a result, across the entire set of guideline items the topic-centered approach scores better (41% against 18%).

Unfortunately, we have observed that for guideline statements with non-zero goal-items, the result sizes of the topic-centric approach are larger. The apparent large difference (1089 against 135.4) is partly caused by an abnormally big size coming from guideline statement 04_6_1. An explanation why this guideline statement leads to such as large result size is that it can find all of the five goal evidences. Thus, the Evidence Coverage Criterion overwhelms the Bounded Number Criteria. When removing this outlier, the difference drops to 287 against 135.4). A similar pattern holds for the overall result sizes.

In the experiment above, we used equal weights for the three criteria, i.e., $k_1 = k_2 = k_3 = 1/3$. The next experiments make a systematic test through different combinations of weights to see which weight values would provide the best results. We consider the following five possible values of the weights to cover the 0-1 interval:

$$\{0, 0.25.0.5, 0.75, 1\}$$

for a single criterion weight.

Table 1: Comparison between topic-centric approach and non-topic-centric approach with $k_1 = k_2 = k_3 = 1/3$.

		Non-topic-centric			Topic-centric		
ID	Original Evidence Number	Found Goal Evidence No.	Found Evidence Number	%	Found Goal Evidence No.	Found Evidence Number	%
04_1_1	5	1	69	20%	2	60	40%
04_1_2	2	1	60	50%	1	166	50%
04_1_3	4	3	327	75%	1	36	25%
04_3_1	4	0	89	0%	0	49	0%
04_3_2	2	2	62	100%	1	28	50%
04_3_3	2	0	27	0%	0	33	0%
04_3_5	2	0	39	0%	1	333	50%
04_3_6	8	3	159	38%	3	140	38%
04_3_7	2	0	52	0%	1	89	50%
04_4_1	5	0	219	0%	3	1628	60%
04_4_2	5	0	42	0%	3	281	60%
04_5_1	3	0	77	0%	0	82	0%
04_6_1	5	0	62	0%	5	9911	100%
04_6_2	3	0	89	0%	1	72	33%
04_7_1	2	0	9	0%	0	372	0%
04_8_1	2	0	15	0%	1	324	50%
Total	56	10	1397	18%	23	13604	41%
No. of non-zero goal evidences		5			12		
Average for non-zero goal evidences			135.4	57%		1089(287)	46%(41%)
Average	3.5		87			850(230)	

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

Because of the normalization condition of the three criteria weights (i.e., $k_1 + k_2 + k_3 = 1$), once two weights are fixed (say, k_1 and k_2), the third weight is also fixed (i.e., $k_3 = 1 - k_1 - k_2$). Thus, the possible combinations of the weights can be considered as a two-dimensional table with $k_1 = 0, 0.25.0.5.0.75, 1$ and $k_2 = 0, 0.25, 0.5, 0.75, 1$ respectively.

Figure 1 shows how many guideline items can find their goal evidence when the weights k_1 and k_2 are set to different values. From the systematic tests with different value combinations, we can see that the system achieves better results when k_2 is set to higher values (i.e., 0.75 or 1). This means that the second criterion (i.e., to check how much of the original evidence which has been used in the current version are covered in the search) plays the most important role on getting better results.

In order to evaluate the proposed approach, we invited three medical professionals from the MAASTRO clinic in the Netherlands to score the guideline update tool with respect to various properties such as functionality, efficiency, usability, reliability and quality of use. The evaluation results are shown in Figure 2, where all the properties are measured on a scale from 1 (worst) to 5 (best).

Component	Update of guidelines		
Evaluator	#1	#2	#3
Functionality	1	1	2
Efficiency	4.33	4.67	3.67
Compatibility	-	-	-
Usability	4.5	3	3
Reliability	5	1	2
Security	?/?	?/?	5
Maintainability	-	-	-
Portability	-	-	-
Quality in use	2.67	2.33	2.67
SUS Score	65	57.5	60

Figure 2: Evaluation by Medical Professional. Scale: 1 (worst) to 5 (best).

The main conclusions from this evaluation by medical professionals are: i) The tool has potential to save time and to identify new relevant evidence for experts who are updating guidelines, ii) There are still too many irrelevant articles suggested as evidence. This produced an overwhelming number of irrelevant articles and reduced the evaluator's overall confidence

in the tool. From that evaluation, we know that the next big step is to improve the precision of the search process [5].

5 CONCLUSION AND FUTURE WORK

In this paper, we have presented a topic-centric approach for searching over new and relevant evidence for updating medical guidelines. We have reported several experiments of the proposed approach and compared our results with those of the non-topic centric approach. The experiments show that the topic centric approach can find goal evidence items for 12 guideline statements out of 16, while the non-topic centric approach can find goal evidence items for only 5 guideline statements. Across the entire corpus of guideline items, the percentage of found goal evidences doubles from 18% to 41%.

Compared with the results of Reinders' approach (Reinders et al., 2015) (with an average result size over one million), the result sizes in our approaches are much smaller. Our approaches are different from Iruetaguena's approach (Iruetaguena et al, 2013), which relies on gathering all relevant articles by searching the PubMed website. Our semantic-distance-based approach can gain a better performance (an average of approximately 10 minutes for each guideline statement) (Hu et al., 2015). There are no differences in the runtime between the non-topic centric approach and the topic centric approach, because adding topic terms in the ranking does not lead to any expensive computation.

There is still future work to improve the existing methods. For example, we can introduce an ontology-based semantic distance measure, so that two semantically equivalent concepts in a medical terminology (says SNOMED CT or UMLS) can be considered to have a zero semantic distance. Thus, relevance measure can be independent from two terms, but instead only depends on the underlying semantic concepts. Another approach to improve the result ranking is to consider the journal classes of the evidence. We can always prefer a publication which appears in a top journal. In future work, we will also do an extensive second evaluation on more medical guidelines.

[5]The software, as well as all experimental data and results is available at http://wasp.cs.vu.nl/sct/download/release/GuidelineUpdateTool-v0.7.zip

ACKNOWLEDGEMENTS

This work is partially supported by the European Commission under the 7th framework programme EURECA Project (FP7-ICT-2011-7, Grant 288048). We thank the clinical trial experts in the MAASTRO clinic for their help on the evaluation.

REFERENCES

Ait-Mokhtar, S., Bruijn, B. D., Hagege, C., and Rupi, P. (2013). Initial prototype for relation identification between concepts, D3.2. Technical report, EURECA Project.

Aït-Mokhtar, S., Chanod, J.-P., and Roux, C. (2002). Robustness beyond shallowness: incremental deep parsing. *Natural Language Engineering*, 8(2):121–144.

Cilibrasi, R. and M.B.Vitanyi, P. (2007). The google similarity distance. *IEEE Trans. Knowledge and Data Engineering*, 19:370–383.

Hu, Q., Huang, Z., den Teije, A., and van Harmelen, F. (2015). Detecting new evidence for evidence-based guidelines using a semantic distance method. In *Proceedings of the 15th Conference on Artificial Intelligence in Medicine(AIME 2015)*.

Hu, Q., Huang, Z., van Harmelen, F., ten Teije, A., and Gu, J. (2014). Evidence-based clinical guidelines in SemanticCT. In *The Semantic Web and Web Science, Volume 480 of the series Communications in Computer and Information Science*, pages 198–212. Springer.

Huang, Z., ten Teije, A., and van Harmelen, F. (2013). SemanticCT: A semantically enabled clinical trial system. In Lenz et al., R., editor, *Process Support and Knowledge Representation in Health Care*. Springer LNAI.

Huang, Z., ten Teije, A., van Harmelen, F., and Ait-Mokhtar, S. (2014). Semantic representation of evidence-based clinical guidelines. In *Proceedings of 6th International Workshop on Knowledge Representation for Health Care (KR4HC'14)*.

Iruetaguena et al, A. (2013). Automatic retrieval of current evidence to support update of bibliography in clinical guidelines. *Expert Sys with Apps*, 40:2081–2091.

NABON (2004). Guideline for the treatment of breast carcinoma 2004. Technical report, Nationaal Borstkanker Overleg Nederland (NABON).

NABON (2012). Breast cancer, dutch guideline, version 2.0. Technical report, Integraal kankercentrum Netherland, Nationaal Borstkanker Overleg Nederland.

NSRS (2006). Guideline complex regional pain syndrome type i. Technical report, Netherlands Society of Rehabilitation Specialists.

Reinders, R., ten Teije, A., and Huang, Z. (2015). Finding evidence for updates in medical guideline. In *Proceedings of HEALTHINF2015*. Lisbon.

Stalnaker, R. (1999). *Context and content*. Oxford: Oxford University Press.

Woolf, S., Grol, R., Hutchinson, A., Eccles, M., and Grimshaw, J. (1999). Clinical guidelines:potential benefits, limitations, and harms of clinical guidelines. *BMJ*, 318(7182):527–530.

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS