Published in final edited form as:

J Intell Inf Syst. 2016 December ; 47(3): 469–490. doi:10.1007/s10844-015-0376-1.

Modelling expertise at different levels of granularity using semantic similarity measures in the context of collaborative knowledge-curation platforms

Hasti Ziaimatin¹, Tudor Groza¹, Tania Tudorache², and Jane Hunter¹

¹School of ITEE, The University of Queensland, Queensland, Australia

²Stanford University, Stanford, CA, USA

Abstract

Collaboration platforms provide a dynamic environment where the content is subject to ongoing evolution through expert contributions. The knowledge embedded in such platforms is not static as it evolves through incremental refinements – or micro-contributions. Such refinements provide vast resources of tacit knowledge and experience. In our previous work, we proposed and evaluated a Semantic and Time-dependent Expertise Profiling (STEP) approach for capturing expertise from micro-contributions. In this paper we extend our investigation to structured micro-contributions that emerge from an ontology engineering environment, such as the one built for developing the International Classification of Diseases (ICD) revision 11. We take advantage of the semantically related nature of these structured micro-contributions to showcase two major aspects: (i) a novel semantic similarity metric, in addition to an approach for creating bottom-up baseline expertise profiles using expertise centroids; and (ii) the application of STEP in this new environment combined with the use of the same semantic similarity measure to both compare STEP against baseline profiles, as well as to investigate the coverage of these baseline profiles by STEP.

Keywords

Expertise profiling; Semantic similarities; Ontologies; Knowledge-curation platforms

1 Introduction

Acquiring and managing expertise profiles represents a major challenge in any organisation, as often, the successful completion of a task depends on finding the most appropriate individual to perform it. Furthermore, the use of expertise profiles to identify, acknowledge and recommend experts from within an online community, motivates additional participants to contribute to the community knowledge base. This collaborative input is vital to the capture and integration of diverse viewpoints and the efficient assembly of an extensive body of knowledge.

Online communities and collaboration platforms have emerged as major components of the Web 2.0 and Semantic Web movements. The old models of how people publish and consume information on the Web have been radically transformed in recent times. Instead of simply viewing information on static Web pages, users now publish their own content through blogs, wikis, photo and video-sharing sites. People are no longer merely consumers of content and applications; they are participants, creating content and interacting with different services and users. More and more people are sharing knowledge through collaborating, discussing and forming online communities; e.g. creating blogs, contributing to knowledge bases such as Wikipedia, where experts share their knowledge and expertise through contributions to the underlying knowledge base. This increase in participation and content creation presents new opportunities for mining expertise from the tacit and implicit knowledge embedded in such platforms.

Collaboration platforms provide a dynamic environment where the content is subject to ongoing evolution through expert contributions. Hence, the knowledge embedded in such platforms is not static as it evolves through incremental refinements to content or through *micro-contributions*. Wiki-based projects, like the the WikiProject Medicine (2014) part of Wikipedia, represent environments where people interested in certain topics (in this case medical and health aspects) collaborate and debate related issues by performing basic editing operations – i.e., creating, editing or removing text and images. For example, the WikiProject Medicine has around 500 members participating in the collaborative authoring of over 31,000 articles. Expert social networks platforms, such as ResearchGate (2014) or Stack Overflow (2014), provide a similar setting by enabling researchers to contribute to Q&A forums. Here, a question or an answer to a posted question can be seen as a *micro-contribution*. A common aspect of both types of platforms is that *micro-contributions* are unstructured – i.e., they are provided as free text entries.

From a different perspective, ontologies have become key elements in the design and development of intelligent decision-support systems, information retrieval systems or knowledge discovery applications and have witnessed an increasing adoption, in particular by the biomedical community (Groza and Tudorache 2013). As a result, ontology engineering has evolved into a community-driven process, where experts focused on a particular domain perform collaborative knowledge curation. Here, instead of contributing and authoring text, experts contribute and author ontological concepts, which due to their intrinsic nature, could be regarded as structured *micro-contributions* to the underlying ontology.

In general, the incremental refinements made by experts in these collaboration platforms provide vast resources of tacit knowledge and experience. Our goal is to tap into this knowledge and create fine-grained and time-dependent expertise profiles. The motivation stems from the increasing trend of relying on online profiles for visibility in particular communities, as well as from the heavy use of these profiles by talent hunters and recruiters.

In a previous publication (Ziainatin et al. 2012), we introduced our Semantic and Time-dependent Expertise Profiling (STEP) approach, for capturing expertise from micro-contributions authored in the context of collaboration platforms that foster knowledge

evolution through incremental refinements. Given a set of micro-contributions, STEP generates short-term and long-term expertise profiles in the form of bags of weighted ontological concepts, using uniformity and persistency as key topic attributes. To date, we evaluated STEP on unstructured micro-contributions – on data retrieved from the Molecular and Cellular Biology (MCB) (2014), and Genetics (2014) Wikipedia projects. The major lesson learned from this evaluation was the need for creating baseline profiles at a level of abstraction closer to the actual micro-contributions. For example, the list of expertise topics manually created by authors of the MCB and Genetics projects included very high-level concepts, such as *Genetics*, *Chemistry*, *Cell* or *Biology*, while the bottom-up profiles (i.e., profiles emerging for direct contributions) included topics, like *Metabolic pathways* or *Lipoprotein lipase*. This makes a direct comparison particularly challenging.

In this paper, we investigate the application of the STEP methodology on structured micro-contributions emerging from the collaborative knowledge engineering effort associated to the International Classification of Diseases (ICD) revision 11 (World Health Organization 2014). ICD is the standard diagnostic classification developed by the World Health Organisation (WHO) to encode information relevant for epidemiology, health management and clinical use. The knowledge curation process of the ICD-11 ontology is done in a collaborative manner by experts from diverse institutions around the world. Each expert contributes to this process by authoring (i.e., creating / modifying / removing) ontological concepts.

We take advantage of the semantically related nature of these structured micro-contributions to showcase two major aspects. Firstly, we propose a novel semantic similarity metric, in addition to an approach for creating bottom-up baseline expertise profiles using expertise centroids – i.e., ontological concepts that act as representatives for an area of the ontology by accumulating high similarity values against all micro-contributions located in that area. These centroids do not only streamline the evaluation methodology, but also provide a more accurate perspective on the actual expertise. Secondly, we describe the application of STEP in this ontology engineering environment and showcase the use of the same semantic similarity metric in comparing resulting profiles to the baseline. Finally, we discuss a method for varying the level of abstraction of the STEP expertise profiles by looking at their coverage of the baseline profiles.

The remainder of the paper is structured as follows. We start with an overview of the STEP methodology and existing work in semantic similarities. Subsequently, we introduce the data used in our experiments and describe the novel semantic similarity metric, in addition to the method for creating expertise centroids. We then present the experimental setup and results and before concluding, we briefly discuss existing expertise modelling / mining solutions.

2 Background

2.1 Semantic and time-dependent expertise profiling model

The STEP methodology consists of three main steps: (i) Concept extraction; (ii) Concept consolidation; and (iii) Profile creation. The *concept extraction* step bridges the gap between the native format of the micro-contributions (e.g., free text in Wikipedia projects or

ontological concepts in ICD-11) and an annotated representation where the micro-contributions are summarised via a set of domain-specific concepts. The actual STEP profile creation process is domain agnostic. Hence, the execution of the Concept Extraction step is delegated to specialised tools that are able to interpret micro-contributions in their raw format (subject to the underlying platform) and to produce an accurate ontological representation. In our previous experiments, Concept Extraction was realised via the NCBO Annotator (Jonquet et al. 2009). In the current context – the ICD-11 ontology – this step is not required, since the micro-contributions have, by default, an ontological format.

The *concept consolidation* step aggregates the diverse manifestations of the resulting ontological concepts and creates *Virtual Concepts* as a semiotical representation of the entities associated with micro-contributions. A *Virtual Concept* represents an abstract entity that clusters semantically related concepts from different ontologies – i.e., concepts that share the same interpretation but which have a slightly different representation because they are not defined in the same ontology.

Finally, the *profile creation* step builds short-term and long-term profiles using consolidated concepts by looking at their uniformity and persistency in the underlying dataset. Both short-term and long-term profiles are sets of weighted concepts, with short-term profiles being restricted to a fixed time window and long-term profiles denoting an overview across all short-term profiles of an expert. In general, expertise profiles created by STEP should be seen as a ranking of the ontological concepts emerging from the micro-contributions. Snapshots of such profiles can be created by applying a desired threshold over the ranking – as shown in our previous evaluation. A complete description of the STEP methodology can be found in Ziaimatin et al. (2012).

2.2 Semantic similarity

Semantic similarities are a critical component of all methodologies dealing with the alignment of resources that can be represented via ontological concepts. For example, an assessment of concept likeness improves the understanding of textual resources and increases the accuracy of knowledge-based applications (Sánchez and Batet 2011). The adoption of ontologies for annotation provides a means to compare entities on aspects that would otherwise not be comparable. For instance, if two gene products are annotated within the same schema, we can compare them by comparing the terms with which they are annotated. While this comparison is often done implicitly (for instance, by finding the common terms in a set of interacting gene products), it is possible to perform an explicit comparison with semantic similarity measures (Pesquita et al. 2009).

In general, a semantic similarity measure is a function that, given two ontology terms or two sets of terms annotating two entities, returns a numerical value reflecting the closeness in meaning between them. Several approaches have been defined for quantifying semantic similarity, the two most prominent ones being: (i) *node-based*, in which the main data sources are the ontological concepts and their properties; and (ii) *edge-based*, which uses the edges between the ontological concepts and their types as the data source. Note that there are other approaches for comparing terms that don't use semantic similarity; for example,

systems that select a group of terms, which best summarise or classify a given subject based on the discrete mathematics of finite partially ordered sets (Pesquita et al. 2009).

Node-based approaches rely on comparing the properties of the terms involved, which can be related to the terms themselves, their ancestors, or their descendants. One concept commonly used in these approaches is Information Content (IC) (Resnik 1995), which provides a measure of how specific and informative a term is. Information Content-based (IC) approaches assess the similarity between concepts as a function of the Information Content share between the concepts. The amount of shared information is represented via the IC of their Least Common Subsumer (LCS) – i.e. the most specific taxonomical ancestor of the two concepts in a given ontology (Resnik 1995). IC quantifies the semantic content of a concept and incorporates taxonomical evidence explicitly modelled in ontologies (such as the number of leaves/hyponyms (specialisations) and ancestors/subsumers. The IC of a concept can be either computed from its probability of occurrence in a corpus (i.e. frequently appearing concepts have lower IC), or from its degree of taxonomical specialisation in the background ontology (i.e. the larger the number of hyponyms (subclasses) of a concept, the more general its meaning and the lower its IC). Pure ontology-based approaches, like the latter one, are preferred to corpora-based ones due to their higher scalability.

Edge-based approaches rely on the structural model defined by the taxonomical relationships in the ontology. These measures base the similarity assessment on the length of the shortest path separating two concepts, defined by going through taxonomical generalisations modelled in the ontology (Rada et al. 1989). The shortest taxonomical path between two concepts is the one that goes through their LCS, which also represents their commonality. Below we list some of the well-known edge-based similarity measures.

Rada (1989), is a simple edge-counting measure, which quantifies the semantic distance of two concepts C_1 and C_2 as the sum of the number of links from C_1 and C_2 to their LCS; i.e. their minimum taxonomical path (1).

$$Dis_{Rada}(C_1, C_2) = N_1 + N_2 \quad (1)$$

Leacock and Chodorow (L&C) (1997) normalise the value by the maximum depth of the taxonomy (D), evaluating the path length in a non-linear fashion (2).

$$Sim_{L\&C} = -\log \left(\frac{N_1 + N_2 + 1}{2D} \right) \quad (2)$$

Wu and Palmer (W&P) (1994) consider the relative depth of the LCS of concept pairs in the taxonomy as an indication of similarity (3).

$$Sim_{w\&p} = -\log \left(\frac{2N_3}{N_1 + N_2 + 2N_3} \right) \quad (3)$$

Other approaches also use path length in addition to other structural characteristics of a taxonomy, such as the relative depth of concepts, and local densities of taxonomical branches. Because several heterogeneous features must be evaluated, these approaches assign weights to balance the contribution of each feature in the final similarity value. These measures, also considered as hybrid approaches, depend on the empirical tuning of weights according to background ontology and input terms, resulting in ad hoc solutions that cannot be easily generalised (Sánchez and Batet 2011).

The main advantage of edge-counting measures is their simplicity. However, edge-based approaches are based on two assumptions that are seldom true in ontologies: (i) nodes and edges are uniformly distributed, and (ii) edges at the same level in the ontology correspond to the same semantic distance between terms. Furthermore, terms at the same depth do not necessarily have the same specificity or semantics, and edges at the same level do not necessarily represent the same semantic distance (Pesquita et al. 2009).

3 Materials and methods

3.1 Data

The target structured micro-contributions have been compiled from the collaborative engineering process of the ICD-11 ontology. The development of ICD-11 is a large-scale project with high visibility and impact for healthcare around the world. The ontology is currently being curated via a shared Web-based process, where many experts contribute, improve and review the domain specific concepts. The collaborative authoring process is similar to any other community curated knowledge bases – e.g., the Molecular and Cellular Biology Wikipedia project – the difference being the resulting data. Here, experts provide incremental changes to ontological concepts, as opposed to free text to an existing article. Hence, the general setting is appropriate for the application of our STEP methodology. In order to have a better understanding of the process, below we provide a brief description of the ICD-11 workflow and data.

A large community of medical experts around the world is involved in the authoring of ICD-11 in a collaborative Web-based platform, called iCAT, a customisation of the generic Web-based ontology editor, WebProtégé (Tudorache et al. 2013b). To date, more than 270 domain experts around the world have used iCAT to author 45,000 classes, to perform more than 260,000 changes and to create more than 17,000 links to external medical terminologies (Tudorache et al. 2013a). iCAT uses the Change and Annotation Ontology (ChAO) (Noy et al. 2006) to represent changes and therefore provides a semantic log of change and annotation data. Change types are ontology classes in ChAO and changes to the ICD-11 ontology are instances of these classes. Similarly, notes that users attach to classes or threaded user discussions are also stored in ChAO. Every change and annotation provides

information about the user who performed it, the involved concept, a timestamp and a short description of the changed or annotated concepts/properties.

We used two main types of data for our analysis: (i) the semantic log of changes and annotations to the ICD-11 ontology; and (ii) the structure of the ICD-11 ontology. The semantic log of changes and annotations to the ICD-11 ontology, extracted from a snapshot of ChAO on 18th March 2014. The following illustrates examples of contributions to ICD-11, extracted from iCAT.

```
URI:    http://who.int/icd#2255_ea0b2e17_d398_4474_8ed0_c2b2ced85d96
Label:  Proliferative Diabetic Retinopathy
Type:   Composite_Change
Date:   05/05/2010 14:53:19
Description: Added a new definition. Prefilled to BC9.2 Proliferative
Diabetic Retinopathy
        Apply to: http://who.int/
icd#2255_ea0b2e17_d398_4474_8ed0_c2b2ced85d96
URI:    http://who.int/icd#1727_ea0b2e17_d398_4474_8ed0_c2b2ced85d96
Label:  Combined arterial and vein occlusion
Type:   Composite_Change
Date:   05/05/2010 13:17:17
Description: Create class with name: H34.81 Combined arterial and vein
occlusion,
        parents: H34.8 Other retinal vascular occlusions
```

ICD-11 has a very large change log; however, the majority of users perform a very small number of changes on a very small number of concepts – i.e., of up to five ontological concepts. The large majority of the changes are due to users that perform bulk operations on a large number of concepts – due to their position in the project; e.g. administrators or group leaders committing or approving a large number of changes. We excluded maintenance changes such as commit and approve which involve a large number of concepts, but do not necessarily reflect the expertise of users performing them. For example, a working group may manage an entire branch of ICD-11, e.g., Infectious diseases, with all changes to its structure and content being committed by a single user at regular intervals. This leads to the user being associated with, for example, 6000+ concepts representing the expertise of the working group, but not necessarily of the user itself.

The filtering process resulted in a total of 19,888 changes by 22 authors involving 737 unique concepts over a period of four years. We were interested in the number of unique concepts to which an expert had contributed (as concepts represent expertise), rather than the total number of changes made by the expert.

The hierarchical structure of the ICD-11 ontology (see an excerpt of the structure in Fig. 1) facilitates access to concepts and their ontological parent-child relationships. We were able

to describe expertise at various levels of abstraction by using semantic similarity measures applied on this structure.

3.2 Semantic similarity measure for creating expertise centroids

A good semantic similarity measure needs to take into account the specific characteristics of the underlying ontology. Our first goal is to define a semantic similarity that accurately reflects the semantics of micro-contributions in close proximity, in addition to their degree of specificity in the larger context of the ontology. We have, hence, adopted a hybrid approach using as foundation the work of Sánchez et al. (2011). More concretely, we redefine an edge-base measure (thus looking at the path between the concepts) in terms of the Information Content (IC) of concepts, which is then expressed via its specificity in the ontology. Contrary to the classical method that computes IC based on term appearance probabilities in a given corpora (Resnik 1995), here, IC is computed using the taxonomic structure of the ontology.

Following the method proposed in Sánchez and Batet (2011), we consider the length of the minimum path separating two micro-contributions (concepts) – i.e., $\min\_path(C_1, C_2)$. This path evaluates the differential semantic features of both concepts as a function of the amount of non-common ancestors found through the shortest link connecting them. In terms of IC, the minimum path length can be approximated as the sum of the amount of differential information between two concepts, as outlined in (4):

$$[IC(C_1) - IC(C_2)] + [IC(C_2) - IC(C_1)] \quad (4)$$

The differential information of one concept compared to another can be quantified by subtracting their common information (i.e. the IC of the LCS of both concepts) from the IC of the concept alone. Formally, this is expressed in (5).

$$\begin{aligned} |\min\_path(C_1, C_2)| &= [IC(C_1) - IC(C_2)] + [IC(C_2) - IC(C_1)] \\ &\cong [IC(C_1) - IC(LCS(C_1, C_2))] + [IC(C_2) - IC(LCS(C_1, C_2))] \\ &= IC(C_1) + IC(C_2) - 2 * IC(LCS(C_1, C_2)) \end{aligned} \quad (5)$$

Subsequently, the depth of a concept C , i.e., the $\min\_path$ between C and the root node, is also redefined in terms of IC, as shown in (6).

$$depth(C) \cong IC(C) + IC(root) - 2 * IC(LCS(C, root)) \quad (6)$$

As the root node is general enough and can potentially subsume any other concepts, its IC could be considered as zero; therefore, the depth of a concept C can be approximated as in (7):

$$depth(C) \cong IC(C) - IC(root) \cong IC(C) \quad (7)$$

Using the definitions above, we redefine the Wu and Palmer similarity measure (3) in terms of IC. Wu and Palmer consider the relative depth of the LCS of concepts in the ontology as an indication of similarity. In other words, in addition to the shortest path separating two concepts, it takes into account the degree of taxonomical specialisation of their LCS. Rada (1989) (1) only considers the length of the minimum path connecting the concepts. Leacock and Chodorow (L&C – (2)) (1997), considers the maximum depth of the ontology; however, in the context of our experiments, this isn't a differentiating factor in determining the similarity between concepts, as all concepts come from the ICD-11 ontology.

Table 1 lists an example of similarity values calculated for two pairs of concepts using the above-mentioned similarity algorithms. The concepts *Enteroviral gastritis* and *Cytomegaloviral gastritis* represent specific types of *Viral gastritis*, a common infection of the stomach and intestines. However, the concepts *Diseases of pancreas* and *Diseases of stomach* represent two different classes of diseases of the digestive system – hence their common LCS, *Diseases of the digestive system*. From a medical perspective, the semantic similarity of concepts in the first pair is higher than the semantic similarity of concepts in the second pair. However, as shown in Table 1, the shortest path between the concepts in both pairs is the same. Furthermore, both Rada and L&C algorithms calculated the same similarity value for concepts in both pairs, despite the fact that the taxonomical specialisation of the LCS of the concept pairs is significantly different, as highlighted by the difference in their depth in the ontology. *Viral gastritis* (depth=8) is more specialised than *Diseases of the digestive system* (depth=4). W&P, on the other hand, calculates a higher similarity between *Enteroviral gastritis* and *Cytomegaloviral gastritis* (similarity=0.889, LCS depth=8) compared to *Diseases of pancreas* and *Diseases of stomach* (similarity=0.8, LCS depth=4). Consequently, we redefined the Wu and Palmer similarity algorithm in terms of IC (8) in order to calculate the pairwise similarity of concepts.

$$\begin{aligned} Sim_{W\&P}(C_1, C_2) &= \frac{2 * depth(LCS(C_1, C_2))}{|min\_path(C_1, C_2)| + 2 * depth(LCS(C_1, C_2))} \\ &\cong \frac{2 * IC(LCS(C_1, C_2))}{IC(C_1) + IC(C_2) - 2 * IC(LCS(C_1, C_2)) + 2 * IC(LCS(C_1, C_2))} \\ &= \frac{2 * IC(LCS(C_1, C_2))}{IC(C_1) + IC(C_2)} \end{aligned} \quad (8)$$

This framework for estimating edge-based similarity measures based on the IC of concepts relies heavily on accurate estimation of IC. In order to calculate the IC of concepts, we analysed a number of approaches, which only considered the subclasses of a concept relative to the maximum number of concepts in the taxonomy, e.g., Seco et al. (2004) and Zhou et al. (2008). None of these approaches consider the depth of a concept as expressed by its subsumers. Consequently, they are unable to differentiate between concepts with the same number of hyponyms/leaves but different depths in the taxonomy. Therefore, we use the approach proposed by Sánchez et al. (2011) which estimates IC intrinsically as the ratio

between the number of leaves of C , as a measure of its generality, and the number of taxonomical subsumers, as a measure of its depth in the ontology (9).

$$IC(C) = -\log p(C) \cong -\log \left(\frac{\frac{|leaves(C)|}{|subsumers(C)|} + 1}{max\_leaves + 1} \right) \quad (9)$$

where $leaves(C)$ is the set of concepts found at the end of the taxonomical tree under concept C and $subsumers(C)$ is the complete set of taxonomical ancestors of C including itself. It is important to note that in case of multiple-inheritance all the ancestors are considered. The ratio is normalised by the least informative concept (i.e., the root of the taxonomy), for which the number of leaves is the total number of leaves in the taxonomy (max_leaves) and the number of subsumers of the root including itself is 1. In order to produce values in the range of 0 and 1 and avoid $\log(0)$, 1 is added to both expressions. This approach also prevents dependence on the specificity and detail of the inner taxonomical structure by relying on taxonomical leaves rather than the complete set of subsumers.

3.3 Creating baseline expertise profiles form expertise centroids

Using the similarity measure defined above, we create baseline expertise profiles by selecting so-called *expertise centroids*. These are concepts associated with micro-contributions that have a high aggregated similarity value across all over micro-contributions found in their close proximity in the ontology. In order to find *expertise centroids*, we compute a matrix of the pair-wise similarity of all concepts to which an expert had contributed, using the measure defined in (8) and described in the previous section.

$$TSim(c, C) = \sum_{i=1}^{n-1} sim(c, ci) \quad (10)$$

Subsequently, for every concept, we calculate the total pair-wise similarity by iterating over all pair-wise similarities computed in conjunction with all other concepts – as per (10). *Expertise centroids*, and the resulting baseline profiles, are then identified by using the median as a threshold over the set of all total pair-wise similarities – as shown in (11).

$$Baseline(C) = \{c | sim(c, C) \geq median[TSim(c_1, C), TSim(c_2, C), \dots, TSim(c_n, C)]\} \quad (11)$$

When calculating the pair-wise similarity of concepts, we used the IC of the LCS of concept pairs; i.e. we used the most taxonomically specific ancestor of concept pairs, to create baseline profiles. However, if we traverse the structure of the ICD-11 ontology and identify ancestors with lower taxonomical specification (lower information content), we can create baseline profiles that contain concepts describing expertise at higher levels of abstraction. In other words, the more taxonomically specific ancestors result in finer-grained profiles, while

ancestors which are less specific and therefore have a lower IC result in profiles containing concepts which represent expertise at higher levels of abstraction.

3.4 Experimental setup

The second goal of our study is to use the baseline expertise profiles to evaluate the STEP methodology on structured micro-contributions. In addition we aim to showcase the possibility of creating expertise profiles at diverse levels of abstraction, by looking at the coverage of the STEP profiles over the baseline. Below we describe the experimental setup for these two aspects.

3.4.1 Evaluating STEP profiles against the baseline expertise profiles—Given two sets of concepts, one representing the STEP profile $SC = \{SC_1, SC_2, \dots, SC_n\}$ and one the baseline $BC = \{BC_1, BC_2, \dots, BC_n\}$, the aim of this task is to find the maximal subset of baseline concepts $\{BC_i \in BC\}$ or the maximal subset of STEP concepts $\{SC_i \in SC\}$ (subject to which initial set is larger) that maximises the overall similarity of SC against BC . To some extent, the underlying principle is the same as in a standard experimental setup in which one requires the computation of Precision / Recall and F-Score, but without relying on exact matching of the candidates against the gold standard. The most important constraint in this setting is that each concept from SC or BC can only be accounted for once – in order to avoid an artificial increase in similarity via multiple counts. The final similarity score is computed as the normalised sum of the pairwise $SC_i - BC_j$ similarity values (based on (8) and (9)), as shown in (12).

$$Similarity(SC, BC) = \operatorname{argmax} \left\{ \frac{p}{n} * \frac{q}{m} * \sum_{i=1}^q sim(c_i, c_k) \right\} \quad (12)$$

where p is the number of concepts matched in BC , $n = |BC|$, $m = |SC|$, q is the number of concepts matched in SC , $c_i \in SC$, $c_k \in BC$ – such that the overall sum is maximised.

In order to have a better understanding of the method, we provide the following example. Let us assume that for a given author, the baseline profile contains concepts C_1 , C_2 and C_3 and the corresponding STEP profile contains concepts C_4 and C_5 . Table 2 illustrates the concept similarity matrix for the profiles, while (13) explains the resulting values.

$$Similarity(STEP, BC) = \frac{2}{3} * \frac{sim(C_1, C_5) + sim(C_2, C_4)}{2} = \frac{2}{3} * \frac{0.98 * 0.52}{2} = 0.47 \quad (13)$$

As illustrated above, due to the single inclusion and maximality constraints, the final matching includes only the concepts C_1 and C_2 from the baseline (because the pair $C_2 - C_4$ has a higher similarity than any of the pairs formed by C_3). Furthermore, we can observe that the maximum overall similarity is achieved by including the pairs $C_2 - C_4$ and $C_1 - C_5$ in the computation, rather than $C_1 - C_4$, even though the similarity of $C_1 - C_4$ (0.73) is

higher than that of $C_2 - C_4$ (0.52). Finally, including (C_3, C_4) or (C_1, C_4) in the overall similarity, would lead to overrepresentation of similarity between the profiles, as the similarity of a single concept in the STEP profile, i.e. C_4 , would be considered with multiple concepts in the baseline profile. In this example, $p = n = 2$ (since all STEP concepts are included in the computation) and $q = 2, m = 3$, since only 2 of the 3 baseline concepts have been used.

To reiterate, the goal is to identify concepts in the compared profiles, which represent similar topics. One extreme comparison, as performed in our previous studies, is to detect exact matches between concepts in the profiles. This leads to underrepresentation of similarity, as different concepts in the two profiles may be semantically similar and represent similar topics. The other extreme is to include the similarity between all concept pairs from the two profiles. This leads to overrepresentation of similarity, as any two concepts will have a similarity value associated with them in the matrix; therefore, we only include the maximum pair-wise similarity of concepts in the overall result, ensuring that the components of every pair are only considered in one pair wise similarity in order to prevent overrepresentation of similarity of the corresponding profiles.

3.4.2 Investigating the coverage of STEP profiles over the baseline expertise profiles—A second aim is to investigate the generation of expertise profiles at different levels of abstraction. The method we have devised to help us perform experiments in this context relies on two aspects:

1. Defining and compiling the subset of baseline concepts that provide a target level of abstraction
2. Defining and computing the coverage of the STEP concepts given the above-defined subset of baseline concepts.

The first of the above aspects has been studied via a clustering approach. More concretely, we have observed the 1:n relationship between each concept in the STEP profile and all combinations of clusters that can be formed from the baseline profile concepts. We quantified this relationship by means of the centrality of the STEP concept in the context of a given cluster – as shown in (14). Centrality is calculated as the normalised sum of pairwise similarity between a STEP concept and a baseline cluster of concepts, normalised by the proportion of baseline concepts in the cluster to the total number of concepts in the baseline profile. The pair-wise similarity uses again the formulations defined in (8) and (9). By measuring this centrality, we are able to understand the extent to which a STEP concept is able to cover (or represent) a set of baseline concepts. And since this relies on a semantic similarity measure that takes into account the path between the concept, as well as the information content, a high centrality score will ensure an appropriate (mid) level of abstraction for the resulting expertise profile.

$$Centrality(c, Cl_n) = \frac{1}{n} * \frac{|Cl_n| - n}{|Cl_n|} * \left[\sum_{i=1}^n sim(c, t_i) \right] \quad (14)$$

where $Cl_n \subseteq Cl$, Cl – the set of all concepts in a baseline profile, $|Cl_n| = n$, $1 \leq n \leq |Cl|$, and $t_j \in Cl_n$.

The final overall coverage is then computed by finding the set of centrality measures that lead to the highest average – (15).

$$Similarity(STEP, Cl) = \frac{1}{|STEP|} * \sum_{i=1}^{|STEP|} \max_i \|Centrality(c_i, Cl)\| \quad (15)$$

where $c_i \in STEP$.

As in the case of the previous section, below we present an example illustrating the computation of this final coverage. We start from $BC = \{C_3, C_4, C_5\}$ and $SC = \{C_1, C_2\}$. The first step is to create all possible clustering combinations from BC – shown in the second column of Table 3. In the second step, we compute the centrality of each concept in SC against each possible clustering option. Assuming that C_1 yields the highest centrality relative to Cl_5 (cluster 5) and C_2 yields the highest centrality relative to Cl_7 (cluster 7), the overall similarity of the STEP and its corresponding baseline profile is obtained by computing the normalised sum of maximum centrality values for the concepts in the STEP profile:

$$Similarity(STEP, BS) = \frac{1}{2} * (Centrality(C_1, Cl_5) + Centrality(C_2, Cl_7)) \quad (16)$$

The best coverage of the baseline profile is determined by considering concepts in the clusters that yield the highest centrality values for concepts in the STEP profile; i.e. cluster 5 and cluster 7. In other words, we take the union of concepts in the clusters that result in the highest centrality for concepts in the STEP profile. In this example, the union of concepts in clusters 5 and 7 is C_3, C_4 and C_5 .

Figure 2 depicts a concrete view over the centrality between a STEP concept (C_1 – *Iron deficiency anaemia due to decreased duodenal absorption*) and a cluster of baseline concepts consisting of C_3 (*Hereditary iron deficiency anaemia*) and C_5 (*Iron deficiency anaemia secondary to blood loss*). As discussed above, the centrality of C_1 is computed in the context of C_3 and C_5 using the semantic similarity measure defined in (8) and (9). In this example, the pair-wise similarity values $sim_{W \& P}(C_1, C_3)$ and $sim_{W \& P}(C_1, C_5)$ are calculated using the same LCS concept, i.e. *Iron Deficiency anaemia*. Furthermore, all concepts have very close IC (9), as they all have the same number of subsumers, and relatively similar number of leaves (C_3 and C_5 have 2 leaves and C_1 has one leaf). This will lead to C_1 having a high centrality value when considered in conjunction with the $C_3 - C_5$ cluster.

4 Experimental results

As mentioned, we have collected structured micro-contributions of 22 experts from the iCat system, each of which had an average of 33.5 ontological concepts they have contributed to. In the initial setup, we created baseline expertise profiles using the proposed semantic similarity measure, as well as STEP profiles by applying STEP directly onto the set of micro-contributions.

We start by analysing the impact of creating baseline profiles. Concepts included in the baseline expertise of an author were selected based on their similarity with other concepts to which the author had contributed. More concretely, a concept has been included in the baseline its total pair-wise similarity with other concepts was greater than the median of total pair-wise similarity of all concepts. It is important to note that using an adjusted median, we can increase the pair-wise similarity threshold to create more abstract profiles or decrease it to create finer-grained profiles.

The results of the baseline profile creation process are presented in Fig. 3. We can observe that the process resulted in a 64.45 % decrease in the number of concepts included in the expertise topic, from an average of 33.5 concepts to 11.91 concepts per author. Qualitatively, the expertise centroids were located, as expected, at a fairly uniform distance (both from a breadth, as well as from a depth perspective) from all over concepts that were in close proximity.

A similar reduction effect can also be observed in the creation of STEP profiles, when taking expertise snapshots by increasing the threshold of the weights associated with the concepts comprised within the profiles. As depicted in Fig. 4, the application STEP methodology on all contributions of an author leads to the inclusion of almost all concepts in the STEP profiles, when no threshold is specified (on average 32.95 concepts per author, compared to the initial 33.5 average concepts per author). A filtering effect is seen when increasing the threshold – initially very restrictive (a reduction of 77.38 % in number of concepts when setting the threshold at 0.05), followed by a quasi-linear behaviour between thresholds 0.05 and 0.15, associated with a decrease from an average of 7.45 concepts / profile at 0.05 to 2.71 concepts / profile at 0.15. Note that STEP profiles are built using a combination between uniformity and persistency. Hence, the increase in threshold leads to retaining only those concepts that are persistent and uniformly distributed throughout the entire time the author has contributed to the project – which is a normal expectation from an expertise profile.

Using the baseline expertise profiles we evaluated the STEP profiles at different thresholds, using the same similarity measure. Experimental results are summarised in Fig. 5 and provided in detail in Fig. 6 – i.e., expanded on all 22 experts.

As it can be observed, when no threshold is imposed, i.e., all concepts in the STEP profiles are included in the comparison, we achieve an almost exact match between the STEP and baseline profiles (99.32 %). Increasing the threshold on STEP to 0.05, 0.1 and 0.15, results in similarities of 49.12 %, 26.17 % and 22.91 %, respectively. The results indicate that with an overall decrease of 77.38 % in the number of concepts (weight threshold = 0.05), we can

achieve 49.12 % similarity against the baseline, while at the highest STEP threshold, where only 8.24 % of concepts are included in the STEP profiles, we achieve an almost 23 % similarity. In practice, this shows that the most restrictive STEP profile is still similar and able to match 23 % of the baseline expertise.

We contrast these results to those discussed in our previous studies (Ziaimatin et al. 2012, 2013), where at weight thresholds of 0, 0.1 and 0.2, we achieved F-score values of 18.91 %, 21.03 % and 20.31 % respectively. While we cannot compare them directly (since the previous results were generated on unstructured contributions and achieved via exact matching), we can draw the conclusion that comparing profiles using semantic similarity methods and the structure of ontologies results in more accurate comparisons than detecting exact matches between the content of profiles. The methods proposed here consider different concepts representing semantically similar topics, while the exact matching method used before considers such concepts as completely different and therefore, results in a less accurate representation of similarity between profiles.

Finally, we have also investigated the extent to which STEP profiles are able to provide a good coverage of the baseline profile, when considering diverse abstraction levels. As shown in the results depicted in Figs. 7 and 8, STEP profiles exhibit an almost constant behaviour in terms of coverage of the baseline profiles, independently of the imposed threshold. The increase in threshold does lead to a small decrease in the centrality of concepts in the STEP profile relative to the optimal subset of baseline concepts. However, this decrease is minimal (on average 2 % per threshold step) and is associated with eliminating the concepts that contribute to noise, rather than excluding concepts in which an author has considerable expertise. This in turn suggests that weights associated with concepts in a STEP profile, represent the true level of an authors expertise in the topics represented by those concepts.

5 Discussion

Our experimental results highlight the significance of semantic similarity methods and the structure of ontologies in profiling expertise at diverse levels of abstraction. However, we have identified a number of limitations of this study, as discussed below:

- Our results are generated using structured contributions in the context of collaborative authoring of the ICD-11 ontology and therefore, should be verified using unstructured contributions, where ontological concepts are derived through annotating experts' micro-contributions, as conducted in our previous research efforts
- We created baseline expertise profiles and used them to evaluate profiles created by the STEP model, however, a comprehensive evaluation can only be performed using a gold standard, which represents the true expertise of an individual with absolute confidence
- The results presented here are based on a snapshot of the ICD-11 ontology, as the ontology is still under development. Hence, in order to create profiles that represent a comprehensive view of experts' expertise, we would need to implement STEP and our proposed methods or fine-tune

the granularity of profiles on the complete set of expert contributions to collaborative authoring of ICD-11.

In this study, every structured contribution identifies the concept which has been the target of the change. Furthermore, all contributions target concepts which belong to the same ontology; i.e. ICD-11. We contrast this to unstructured contributions in the context of collaboration platforms such as MCB or Genetics Wiki projects, where we extract contributions in natural language form and annotate them in order to map identified expertise topics to ontological concepts. As various domain ontologies are used for annotating contributions, identified expertise topics are often mapped to multiple concepts from different ontologies.

In order to apply these methods to unstructured contributions, in the future, we plan to leverage *ontological lenses*. An *ontology lens* provides a domain-specific view over the expertise of an individual by considering concepts that emerge from the annotation of an expert's contributions using a given ontology; e.g. all SNOMED-CT concepts that have emerged from annotating an expert's contributions, will constitute a lens; or the GO lens will contain all concepts that emerge from annotating an expert's contributions using the Gene Ontology (GO). We subsequently identify the ontology lens that best describes the expertise of the expert – i.e., the one that contains the highest number of concepts. We will then use the structure of the corresponding ontology to apply the semantic similarity methods we have proposed in this paper.

6 Related work

Expertise profiling is an active research topic in a wide variety of applications and domains, including biomedical, scientific and education. In this section, we present a brief overview of the related efforts, with particular emphasis on Information Retrieval (IR) and Semantic Web domains. To the best of our knowledge, no research has specifically focused on creating semantic and time-dependent expertise profiles using micro-contributions in the context of collaboration platforms where the content is dynamic and continuously evolves – as in this particular case, the iCAT platform. All existing expertise profiling / matching systems analyse large corpus of static documents such as researcher publications, grants, awards, etc in order to profile the expertise of an individual. Below, we present a brief overview of such approaches.

Most research efforts focus on the task of expert finding and propose methods for finding experts, given a query or knowledge area in which experts are sought. Not only is expert finding a different task to expert profiling, but the methods applied in such studies rely on large corpus of static documents (e.g., publications) and therefore are not suitable in the context of shorter text, such as micro-contributions in the context of living and evolving documents. The Entity and Association Retrieval System (EARS) (Balog 2008) is an open source toolkit for entity-oriented search and discovery in large test collections. EARS, implements a generative probabilistic modelling framework for capturing associations between entities and topics. Currently, EARS supports two main tasks: finding entities (“*which entities are associated with topic X?*”) and profiling entities (“*what topics is an*

entity associated with?). EARS employs two main families of models, both based on generative language modelling techniques, for calculating the probability of a query topic (q) being associated with an entity (e), $P(q|e)$. According to one family of models (Model 1) it builds a textual representation (i.e., language model) for each entity, according to the documents associated with that entity. From this representation, it then estimates the probability of the query topic given the entity's language model. In the second group of models (Model 2), it first identifies important documents for a given topic, and then determines which entities are most closely associated with these documents. In previous studies, we conducted experiments with EARS using our biomedical use cases (see Ziaimatin et al. 2012, 2013); however, this system also relies on a given set of queries. Furthermore, as with other studies that target expert finding, EARS relies on large corpus of static publications, while we aim at building expert profiles from micro-contributions, without relying on any queries.

In the same category of expertise finding, we find SubSift (short for submission sifting), which is a family of RESTful Web services for profiling and matching text (Price et al. 2013). It was originally designed to match submitted conference or journal papers to potential peer reviewers, based on the similarity between the papers' abstracts and the reviewers' publications as found in online bibliographic databases. In this context, the software has already been used to support several major data mining conferences. SubSift, similar to the approaches discussed above, relies on significant amounts of data and uses traditional IR techniques such as TF-IDF, Bag-of-Words and Vector based modelling to profile and compare collections of documents.

The Saffron system provides users with a personalised view of the most important expertise topics, researchers and publications, by combining structured data from various sources on the Web with information extracted from unstructured documents using Natural Language Processing techniques (Monaghan et al. 2010). It uses the Semantic Web Dog Food (SWDF) (Möller et al. 2007) corpus to rank expertise and makes a distinction between the frequency of an expertise topic occurring in the context of a skill type and the overall occurrence of an expertise topic. Saffron also extends information about people by crawling Linked Open Data (LOD) (Bizer et al. 2009) from seed URLs in SWDF. The semantics of the SWDF and crawled data represented using Semantic Web technologies is consolidated to build a holistic view represented via the social graph of an expert.

Existing social networks such as BiomedExperts (BME) (2014) provide a source for inferring implicit relationships between concepts of the expertise profiles by analysing relationships between researchers; i.e., co-authorship. BME is the world's first pre-populated scientific social network for life science researchers. It gathers data from PubMed on authors' names and affiliations and uses that data to create publication and research profiles for each author. It builds conceptual profiles of text, called Fingerprints, from documents, Websites, emails and other digitised content and matches them with a comprehensive list of pre-defined fingerprinted concepts to make research results more relevant and efficient.

Similarly, Profiles Research Networking Software (2014), is an NIH-funded open source tool to speed the process of finding researchers with specific areas of expertise for

collaboration and professional networking. Profiles RNS imports and analyses “*white pages*” information, publications and other data sources, to create and maintain a complete search-able library of web-based electronic CVs. Profiles RNS self-populates a database of publication history, research interests and professional relationships for each investigator in an organisation.

Finally, Pure (2014), an advanced enterprise-wide system, enables expertise profile management by aggregating an organisations research information from numerous internal and external sources. It analyses a researcher’s publications and awards and transforms them into a unique Fingerprint, a distinct visual index of concepts and a weighted list of structured terms.

7 Conclusions and future work

In our previous work, we applied the STEP model to unstructured contributions in the context of the MCB and Genetics Wikipedia projects and created fine-grained and time-dependent expertise profiles. Our baseline profiles consisted of high-level description of expertise provided by corresponding authors upon joining the projects. Hence, we were unable to compare STEP and baseline profiles directly, as they represented expertise at significantly different levels of abstraction. In this study, we: (i) analysed the performance of the STEP model in the context of structured contributions to collaborative authoring of the ICD-11 ontology; (ii) applied semantic similarity measures to micro-contributions in order to create fine-grained baseline profiles at a level of abstraction comparable with the STEP profiles; and (iii) investigated the alignment and coverage between STEP and baseline profiles.

Future work will focus on creating *ontology lenses* on profiles built by STEP. This provides the opportunity to view a long-term profile from different ontological perspectives, each of which only considers concepts from a particular ontology. From an abstract perspective, since a given ontology represents the conceptualisation of a specific domain, ontology lenses represent a domain-specific view over the expertise of an individual. Using *ontology lenses*, we can also determine the accuracy with which a given ontology represents the expertise of an individual. More importantly, ontology lenses will facilitate the application of similarity measures for creating profiles with various degrees of abstraction, in platforms with unstructured contributions.

Acknowledgments

This research is partly funded by the Australian Research Council (ARC) Discovery Early Career Researcher Award (DECRA) – DE120100508.

References

1. Balog, K. PhD thesis. University of Amsterdam; 2008. People search in the enterprise.
2. [Accessed 17 September 2014] Biomed Experts. 2014. <http://www.biomedexperts.com/>
3. Bizer C, Heath T, Berners-Lee T. Linked data – the story so far. International Journal on Semantic Web and Information Systems. 2009; 5(3):1–22.

4. Chodorow M, Leacock C. Combining local context and WordNet similarity for word sense identification. *Fellbaum*. 1997:265–283.
5. Groza T, Tudorache T, Dumontier M. State of the art and open challenges in community-driven knowledge curation. *Journal of Biomedical Informatics*. 2013; 46(1):1–4. [PubMed: 23219718]
6. Jonquet C, Shah NH, Musen MA. The open biomedical annotator. *Summit on translational bioinformatics*. 2009; 2009:56–60. [PubMed: 21347171]
7. Möller, K.; Heath, T.; Handschuh, S.; Domingue, J. Recipes for semantic web dog food: The ESWC and ISWC metadata projects. *Proceedings of the 6th International Semantic Web Conference*; Springer; 2007. p. 802-815.
8. Monaghan, F.; Bordea, G.; Samp, K.; Buitelaar, P. Exploring your research: Sprinkling some saffron on semantic web dog food. *Proceedings of the semantic web challenge at the international semantic web conference*; 2010.
9. Noy, NF.; Chugh, A.; Liu, W.; Musen, MA. A framework for ontology evolution in collaborative environments. *Proceedings of the 5th international semantic web conference*; Springer; 2006. p. 544-558.
10. Pesquita C, Faria D, Falcao A, Lord P, Couto F. Semantic similarity in biomedical ontologies. *PLoS Computational Biology*. 2009; 5(7)
11. Price S, Flach PA, Spiegler S, Bailey C, Rogers N. SubSift web services and workflows for profiling and comparing scientists and their published works. *Future Generation Computer Systems*. 2013; 29(2):569–581.
12. [Accessed 17 September 2014] Professional Networking and Expertise Mining for Research Collaboration. 2014. <http://profiles.catalyst.harvard.edu/>
13. [Accessed 17 September 2014] Pure. 2014. <http://info.scival.com/pure>
14. Rada R, Mili H, Bicknell E, Blettner M. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man and Cybernetics*. 1989; 19(1):17–30.
15. Resnik, P. Using information content to evaluate semantic similarity in a taxonomy. *Proceedings of the 14th IJCAI*; 1995. p. 448-453.
16. [Accessed 17 September 2014] ResearchGate. 2014. <http://www.researchgate.net/>
17. Sánchez D, Batet M. Semantic similarity estimation in the biomedical domain: An ontology- based information-theoretic perspective. *Journal of biomedical informatics*. 2011; 44(5):749–759. [PubMed: 21463704]
18. Seco N, Veale T, Hayes J. An intrinsic information content measure for semantic similarity in WordNet. *Proceedings of ECAI 2004*. 2004:1089–1090.
19. [Accessed 17 September 2014] Stack Overflow. 2014. <http://stackoverflow.com/>
20. Tudorache, T.; Nyulas, C.; Noy, NF.; Musen, MA. Using semantic web in ICD-11: three years down the road. *Proceedings of the 12th international semantic web conference*; Springer; 2013a. p. 195-211.
21. Tudorache T, Nyulas C, Noy NF, Musen MA. WebProtégé: A collaborative ontology editor and knowledge acquisition tool for the Web. *Semantic Web Journal*. 2013b; 4(1):89–99.
22. Wikipedia. [Accessed 17 September 2014] WikiProject Genetics. 2014. <http://en.wikipedia.org/wiki/Wikipedia:WikiProjectGenetics>
23. Wikipedia. [Accessed 17 September 2014] WikiProject Medicine. 2014. <http://en.wikipedia.org/wiki/Wikipedia:WikiProjectMedicine>
24. Wikipedia. [Accessed 17 September 2014] WikiProject Molecular and Cellular Biology. 2014. <http://en.wikipedia.org/wiki/Wikipedia:MCB>
25. World Health Organization. [Accessed 17 September 2014] International Classification of Diseases. 2014. <http://www.who.int/classifications/icd/en/>
26. Wu, Z.; Palmer, M. Verb semantics and lexicon selection. *Proceedings of the 32nd ACL*; 1994. p. 133-138.
27. Zhou, Z.; Wang, Y.; Gu, J. A new model of information content for semantic similarity in WordNet. *Proceedings of the 2nd international conference on future generation communication and networking symposia*; IEEE; 2008. p. 85-89.

28. Ziaimatin H, Groza T, Bordea G, Buitelaar P, Hunter J. Expertise profiling in evolving knowledge-curation platforms. *GSTF Journal on Computing*. 2012; 2(3):118–128.
29. Ziaimatin H, Groza T, Hunter J. Semantic and time-dependent expertise profiling models in community-driven knowledge curation platforms. *Future Internet*. 2013; 5(4):490–514.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Fig. 1.

Excerpt from the ICD-11 Ontology showing its high-level structure. The classification of diseases in ICD-11 starts with a set of well-defined branches (e.g., Infectious diseases, Diseases of the circulatory system, etc) and is refined into sub-categories, with the leaves of the ontology representing instances of actual diseases

Fig. 2.
Excerpt from the ICD-11 Ontology used to exemplify the computation of the coverage of
STEP expertise profiles

Fig. 3.

The creation of baseline expertise profiles from the total number of concepts authored by each of the 22 experts leads a 64.45 % decrease in the number of concepts, from an average of 33.5 concepts to 11.91 concepts per author

Fig. 4.

The effect of varying the weight threshold over STEP profiles. When no threshold is specified, the resulting profile mirrors the set of initial micro-contributions – on average 32.95 concepts per author, compared to the initial 33.5 average concepts per author. A filtering effect is seen when increasing the threshold – initially very restrictive (a reduction of 77.38 % in number of concepts when setting the threshold at 0.05), followed by a quasi-linear behaviour between thresholds 0.05 and 0.15, associated with a decrease from an average of 7.45 concepts / profile at 0.05 to 2.71 concepts / profile at 0.15

Fig. 5. Summarised representation of the evaluation of STEP profiles using the baseline expertise profiles

Fig. 6. Expanded representation of the evaluation of STEP profiles using the baseline expertise profiles

Fig. 7.
Summarised representation of investigating the baseline converge of STEP profile

Fig. 8.
Expanded representation of investigating the baseline converge of STEP profiles

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

An example of concept similarity calculated for two pairs of concepts using various algorithms

Table 1

Concept pair	LCS	LCD depth	Path	Rada	L&C	W&P
Enteroviral gastritis	Viral gastritis	8	2	2	2.30	0.889
Cytomegaloviral gastritis						
Diseases of pancreas	Diseases of the digestive system	4	2	2	2.30	0.8
Diseases of stomach						

Table 2

Example of the similarity matrix computed for comparing a STEP profile to a baseline profile

STEP concepts	<u>Baseline concepts</u>		
	C_1	C_2	C_3
C_4	0.73	0.52	0.31
C_5	0.89	0.01	0.24

Table 3

Example of the similarity matrix for a STEP and its corresponding baseline profile

Cluster	STEP concept	Centrality
Cl_1	C_3	$Centrality(C_1) = sim_{W \& P}(C_1, C_3)$ $Centrality(C_2) = sim_{W \& P}(C_2, C_3)$
Cl_2	C_4	$Centrality(C_1) = sim_{W \& P}(C_1, C_4)$ $Centrality(C_2) = sim_{W \& P}(C_2, C_4)$
Cl_3	C_5	$Centrality(C_1) = sim_{W \& P}(C_1, C_5)$ $Centrality(C_2) = sim_{W \& P}(C_2, C_5)$
Cl_4	C_3, C_4	$Centrality(C_1) = 2/3 * 1/2 * [sim_{W \& P}(C_1, C_3) + sim_{W \& P}(C_1, C_4)]$ $Centrality(C_2) = 2/3 * 1/2 * [sim_{W \& P}(C_2, C_3) + sim_{W \& P}(C_2, C_4)]$
Cl_5	C_3, C_5	$Centrality(C_1) = 2/3 * 1/2 * [sim_{W \& P}(C_1, C_3) + sim_{W \& P}(C_1, C_5)]$ $Centrality(C_2) = 2/3 * 1/2 * [sim_{W \& P}(C_2, C_3) + sim_{W \& P}(C_2, C_5)]$
Cl_6	C_4, C_5	$Centrality(C_1) = 2/3 * 1/2 * [sim_{W \& P}(C_1, C_4) + sim_{W \& P}(C_1, C_5)]$ $Centrality(C_2) = 2/3 * 1/2 * [sim_{W \& P}(C_2, C_4) + sim_{W \& P}(C_2, C_5)]$
Cl_7	C_3, C_4, C_5	$Centrality(C_1) = 3/3 * 1/3 * [sim_{W \& P}(C_1, C_3) + sim_{W \& P}(C_1, C_4) + sim_{W \& P}(C_1, C_5)]$ $Centrality(C_2) = 2/3 * 1/2 * [sim_{W \& P}(C_2, C_3) + sim_{W \& P}(C_2, C_4) + sim_{W \& P}(C_2, C_5)]$