Proceedings of the International Conference on Semantic Web Business and Innovation (SWBI2015), Sierre, Switzerland, 2015

Ontology-based Recommendation from Natural Language User Descriptions: Recommending Eco-Innovations to Companies

Csaba Oravecz Bálint Sass Csongor Sárközy
Research Institute for Linguistics
Benczúr u. 33. 1068 Budapest, Hungary
{oravecz.csaba,sass.balint,sarkozy.csongor}@nytud.mta.hu

ABSTRACT

The paper reports on the development of a recommendation system which offers complex services and green technology solutions for companies. The system must work in a predominantly data sparse environment with respect to traditional evidence of user preference and therefore has to seek other ways of collecting the information necessary for principled recommendations. We utilize and adapt methods from natural language processing, text analysis, information retrieval, information extraction and knowledge engineering in order to overcome the data sparseness problem. The system works with Hungarian language data but with the necessary resources developed the framework is applicable in other languages as well.

KEYWORDS

domain ontology, knowledge based recommendation, profile vector similarity, information extraction, green technology

1 INTRODUCTION

Using recommender systems has become practically everyday practice in matching supply and demand by giving assistance to users in reaching the most interesting items in online retrieval environments. In the typical scenario a prediction is computed about the relevance of the available items users have not seen or are aware of, taking into account their profile which can be calculated in various ways. Based on the information these systems make use of and the methods by which the prediction is calculated two basic recommendation strategies are normally distinguished, content-based and collaborative filtering types [1]. Both have their strengths and weaknesses and it is common to combine the two approaches into hybrid systems for better performance with the possibility of integrating further knowledge sources leading to the now very popular semantically enhanced models exploiting complex semantic relations [2–4].

A standard property of these systems is that they operate in an environment where there is significant amount of data available, typically in some online marketplace with a huge number of various kinds of items and a large population of users, and systems collect and aggregate this data and use it when running the appropriate recommendation algorithms. It is therefore expected that in low data density situations new difficulties and challenges may arise that need to be resolved. In this paper we describe the development of a system that has to handle several parameters that can be considered uncommon in a typical recommendation situation: (i) the domain environment, (ii) the recommended items and (iii) available user information.

1.1 Motivation

New green and sustainable solutions appear continuously in many domains, including energetics, architecture, waste management and recycling, material assessment, transportation, logistics and others, many of them may help users (companies) not only in sustainable and responsible business development abut also in finding new economic ways to reduce expenditures. Sustainable solutions through green technologies to daily problems offer great business opportunities.

The objective of our project was to set up a user-friendly online platform for green technology and knowledge transfer to Hungar-

ian companies, supporting environmentally friendly technologies in the industrial production of the region by providing information on eco-innovations, including technologies, applications, products, processes and other solutions. The standard practice in this field is using high level of human expertise to carry out a time consuming and labour intensive audit process for the companies identifying the areas where these solutions could be beneficial. However, manual auditing of companies is a prohibitively resource demanding task, so our system aims to replace this process by recommending and so promoting eco-innovations (any new product, service or process that benefits the environment) to market participants in a domain that has enormous reserves both on the supply and demand side, but connecting them is encumbered by a huge knowledge gap. It features eco-innovative results relevant to all economic sectors and in different implementation stages. These can range from knowledge, guidelines, processes, products and applications, patents as well as eco-innovation networks and other information sources. Using a recommender supports clients without extensive domain knowledge in finding relevant results, which are delivered to those who may lack the time to actively look out for innovative input for themselves.

1.2 Challenges

The system has to work in a domain of application where data is on the one hand fairly complex and on the other hand very sparse, so the necessary information might not be readily available from the traditional sources. Using the standard vocabulary of recommendation our *items* are complex services and solutions whose relevant properties and their representation are not trivial to specify. However, the amount of these items is several magnitude lower[1] than those (books, movies, documents etc) in a classic recommendation scenario, which enables a precise manual annotation for item representation in the system.

[1] A few hundred items at most.

Practically all *users* (companies, enterprises etc.) of the system are first time clients and their limited interaction will typically never accumulate the amount of input data to build a reliable profile[2] from, resulting in perhaps the worst (and persistent) *cold-start problem* ever. Instead, information about items and users are available from alternative sources:

- For items: unstructured or semistructured natural language descriptions of services and solutions not primarily intended for computational processing.

- For users: all harvestable information from web sites, public documents, company brochures, reports etc. available electronically either in a passive way (automatically crawled by the system upon registration and providing the URL(s)) or actively by user upload. These also have almost exclusively free text content.

Furthermore, users are not expected to have any knowledge of the domain and items of recommendation and by default cannot be required to provide any item specific assistance to the system (such as specifying their preferences with respect to item properties).

2 THE OVERALL RECOMMENDATION MODEL

The objective was to develop a system with the capability of adaptation to the initial conditions and utilising all available information while reducing the required active user input to the minimum.[3] Combining standard and well-known techniques from various fields in an efficient way the system offers a robust, flexible and tunable solution to the above challenges. The specific and constrained domain allows for a detailed representation of relevant information, background knowledge and concepts in a common space of an OWL ontology. Within

[2] Called *behaviour-based* profile in [5].

[3] The complex interaction that a self-assessment for a company of their eco-energy performance requires was ruled out from the beginning.

ISBN: 978-1-941968-19-2 ©2015 SDIWC

the space of this knowledge base the recommendation task can be converted into a classic information retrieval problem: after a proper representation for the recommended items (as "documents") as well as for the relevant properties of users (as "queries") is computed, recommendation will be reduced to the calculation and ranking of the similarities between these representations. Given the formal specification of the domain knowledge in the ontology, an obvious candidate for the representation of relevant entities is a (binary or real) vector whose coordinates are the concepts in the ontology [6]. Similarities can then be calculated straightforwardly using standard vector space measures like the cosine similarity [7, 8].

3 RELATED WORK

In this section we summarise the approaches our work is most closely related to, highlighting the similarities and the small but important differences as we go along. Since we draw upon several common methods of a wide range of fields (from information retrieval to ontology engineering) the related literature is huge and many components of our approach will appear in several different sources. Therefore our selection here cannot be exhaustive and remains in many respects only exemplary.

Due to the lack or limited amount of *behaviour-based* data, the constrained and well definable domain and the "manageable" number of items to be recommended, our system crucially depends on an extensive and detailed knowledge base encoded in the form of an ontology forming the conceptual backbone and common semantic space for the item and user profiles. *Semantics* and *ontology* have become buzzwords in many fields of information management and recommendation is no exception. There is a wide range of research from [9] to [10] focusing on the semantic enhancements of recommenders by using ontologies as a basis for storing the relevant knowledge, and inferring relations. A nice concise summary of semantically enhanced recommendation models is given in [4] where they divide the recommendation process into three main steps: (i) domain knowledge acquisition and semantic analysis of content, (ii) concept-based entity modelling and semantic profile expansion, (iii) semantic-matching strategy for prediction; and classify the related research according to the methods and algorithms they use during these steps. Our approach is basically a selection of these methods adjusted to the circumstances.

There is an important distinction how ontologies are used in recommendation systems. In many cases they actually store or include the items to be suggested as instances [11–13] and consequently user profile is usually represented on the basis of items and their related concepts selected by the user in earlier interaction with the system[4]. In our system, however, it is the relevant domain that is captured by the knowledge base (as for example in [14]), and items are not in themselves are stored as instances in this KB. Specifically, users and items are represented in terms of how they can be linked to relevant parts of the domain.

Modelling the entities in the common semantic space requires the extraction and/or transformation of the information from the available sources into the selected representations. Our sources amenable to automatic processing are natural language texts for which there is a long tradition of semantic information extraction. To take just a few examples, KIM [15] applies some basic NLP methods like named entity recognition to carry out the semantic (or entity) annotation, using the instance labels of the knowledge base to find potential occurrences of concepts in text documents. [16] is in line with the popular approaches of recognising some patterns of various complexity over different levels of abstraction in data[5], and relies primarily on regular expressions. Cerno is claimed to be a "light-weight framework for semantic annotation of textual documents using domain-specific ontologies" [17], combining keyword and structure-based annotation

[4] The default profile can then be expanded with different strategies.

[5] By this we mean first the actual data (strings of characters as the raw text) and then any kind of annotation based on this data, such as segmentation or some linguistic labelling.

ISBN: 978-1-941968-19-2 ©2015 SDIWC

rules instead of linguistic patterns. For the construction of our knowledge base we use a similar set of steps they applied to develop the domain dependent vocabulary for their annotation schema, this will be discussed in more detail in Section 4.1 They also define a set of common requirements against a semantic annotation system, such as adaptability, portability, accuracy and efficiency, and scalability, which are to be fulfilled by our system as well, with one important difference: they aim for quick but not necessarily very accurate analysis of documents while we are ready to settle for less speed but need high accuracy.

[18] is a good summary of extracting semantic information from text using ontologies and linguistic templates. We second their claim that for high recall and precision hand-crafted methods are superior for information extraction in a rigid and structured domain and follow their approach in using hand-crafted templates. We note here that we also take a similar route to theirs in the iterative development process in the ontology engineering phase, starting with a core ontology including the basic concepts and a simple hierarchy, then during repeated experiments with this ontology in reasoning and searching the base is extended for best performance (see Section 4.1). Their information extraction module, however, does not use linguistic tools such as part-of-speech taggers, parsers or phrase chunkers; only a named entity recogniser and a lexical analyser are applied, and the complex semantic entities and relations are extracted according to the predefined templates.

Once the relevant default profiles are computed in the semantic space, for the full exploitation of the power of the knowledge base it is necessary to discover as much of the available semantic relations as possible, i.e. to enhance the profile representations. Two standard methods are normally applicable [19]:

- Employing (constrained) spreading activation (CSA) [20] in the semantic network, where the scores of a set of concepts are propagated to other related concepts through the concept relations, with many possible different parameter settings (weights) and ways of placing constraints on the spreading of the activation.
- Applying complex domain-based inferential processes based on the internal structure and relations defined by the ontology. A wide range of measures are used to calculate or weight these inferences, whose only real test can be empirical evaluation as opposed to some independent logical justification.

The latter approach is used in for instance [11] and [21], where they discover semantic associations along property sequences in which the semantic intensity of nodes is calculated from several component similarity measures.

The former approach is more straightforward and has widespread application in related work. Since we also use this method for profile extension, we give a brief survey of a few more examples from related research. [22] uses CSA in a semantic domain model to improve web search. In [23] user interests are presented as keyword vectors and spreading activation helps to build user profiles constructed from past usage behaviour. They distinguish three major strategies for spreading that are commonly used in CSA: generalisation, specialisation, and relevance expansion. Generalisation is activating a concept above the current concept, specialisation is that below the current concept[6]. By relevance expansion they refer to the process of activating a concept through a non-*is-a* relationship in the network.

[24] also encode user preferences and item features in terms of semantic concepts defined in domain ontologies representing the domain of discourse where user interests are defined. The semantic preference spreading mechanism expands the initial set of preferences stored in user profiles through explicit semantic relations with other concepts in the ontology. They control the expansion by applying a decay factor to the intensity of preference each time a

[6]This is not used in our system since linking to a more general concept does not entail that the profile can also be reliably related to the more specific concepts.

relation is traversed. They also argue that basic user profiles without expansion are too simple to deliver good results.

[13] devise an extended spreading activation technique (ESAT) for ontology based user profile extension. Their main focus is the personalisation of user profile and they argue for the limitation of the basic SA algorithm from the point of view of personalised ontological profiles. It is well out of the scope of the present paper to go into details but the objections they raise seem arguable[7], and their evaluation shows that user profile personalisation, i.e. grouping related concepts together[8] is most effective if rich enough user ratings are available, otherwise spreading activation alone gives good performance. According to standard practice, they attach weights to properties (relations) then compute semantic relatedness using these weight values. We note here that there could be a multitude of methods devised to calculate relatedness within an ontology and probably the best way of selecting among them is through extensive empirical evaluation. To make a shortcut however, in our approach this step is simply taken over by the spreading algorithm. A lot of the methods to work out these semantic relations seem fairly speculative and so we would like to show that the a system can work well without recourse to these complex calculations. This does not mean that our framework is not open and cannot be extended by application of these techniques presented for example in [13].

Recommendation is implemented by calculating the similarities between the extended item and user profiles. Note that with the representation of these profiles as vectors in the common semantic space, the coordinates being the concepts of the full ontology with the appropriate values filled in, it is no longer true that a standard vector similarity measure can only detect "similarity when the considered item has exactly the same features defined in the user profile, thus preventing any semantic reasoning process" [11], since all profile vectors are in the same space and of necessity will have the same dimensions. Similarity is then easily calculated by the cosine function, which is used by many authors in several domains including [8], [7] or [25].

In summary we can consider [26] and [14] as the closest approaches to ours in using similar algorithms and system architecture. The former is of course not about the same task and furthermore does not address the problem of knowledge extraction from text but only provide a vocabulary and some simple mechanisms to aid in the semi-automatic annotation of documents. In the latter they rely only on names of classes and instances to construct concept vectors while we apply a purpose-built separate mapping layer in the form of the lexical templates (see Section 4.3.1) to assign ontology concepts to documents, and they do this with recommended items and we do it with user data as well, but eventually the main components draw upon similar methods.

4 SYSTEM ARCHITECTURE

In this section we describe in some detail the main components of the system and the recommendation process that builds upon these components.

4.1 Creating the knowledge base

All of the background knowledge about the application domain, relevant objects, categories, concepts and relations are encoded in an explicit formal knowledge base implemented as an OWL ontology. For the development we have been using the systematic process similar to the common requirements in knowledge engineering approaches [17]:

- Identification and collection of concepts by the help of domain experts.

[7] One limitation they would like to point out is that SA considers just the main structure of a network (the reference ontology) but not the structure of user ontological profiles. It is clear however, that if SA starts from the nodes activated by the user profile (i.e. nodes having non-zero values in the user's profile vector) user preference will necessarily be taken into account.

[8] One can devise other 'hidden' semantic relations to uncover so it is not the only possibility.

- Extension of the base set of concepts with background knowledge and related concepts from using available general and domain specific knowledge sources.
- Structuring the collected information in a formal semantic model.

```
1993.55969884557  1356 hulladék
                       ("waste")
1553.50432823931  1087 környezeti
                       ("environmental")
1529.70377199569  1104 megtakarítás
                       ("reduction of
                         expenses")
1385.73629885459  1115 intézkedés
                       ("measure")
1068.88348017766   786 tonna
                       ("ton")
...
```

Figure 1. Keyword list from domain specific documents.

```
nyers<>fűrészpor<>12.2717 3.1616
("raw sawdust")
kompakt<>fénycső<>11.7500 3.4631
("compact fluorescent")
szabadlevegős<>hűtés<>11.7395 3.4631
("open-air cooling")
szerves<>oldószer<>10.7642 3.8708
("organic solvent")
mart<>aszfalt<>12.7571 3.1618
("milled asphalt")
hulladékhő<>hasznosítás<>9.9866 3.1592
("waste heat recovery")
épület<>fűtés<>8.5454 3.4548
("building heating")
...
```

Figure 2. Multiword expressions with association measure scores from domain specific documents.

The identification and generation of the domain dependent set of concepts was aided by standard NLP methods for keyword and phrase extraction [27, 28] from a collection of domain specific documents[9], resulting in large lists of linguistic resources (see Figure 1 and 2 for illustration), which were then manually filtered and cleaned up to be converted into concept nodes in the ontology. Additional relevant concepts were obtained from available lexical resources such as the Hungarian WordNet [29], and predefined taxonomies like the European Waste Catalogue (EWC) and the Hungarian NACE, the classification of economic activities (HuNACE).[10]

Clearly, knowledge base development is a never ending enterprise and an ideal state of affairs can never be reached even in a restricted environment. Unfortunately, in the application domain of our system this restriction applies only in the set of items (the supply side) but hardly in user activities (demand side), since potential clients, companies can deal with practically anything and there are very general solutions (like isolation of buildings or innovative heating solutions) that are applicable for a wide range of activities: there is no clear limitation or focus given to an identifiable group of SMEs, sector, branch, size or business form, all can benefit from eco-innovation results and solutions. It is simply impossible to cater for such a potentially broad range of topics in the knowledge base and therefore it is essential to focus on selected areas. In our current system it is waste management that is encoded and worked out in most detail as a high impact area in green technologies. A fragment of our current ontology, which at current stage contains altogether about 2500 concepts, is illustrated in Figure 3. The ontology, together with the other resources, is constantly going through an iterative development process during repeated experiments and testing with the system.

Now let us assume that the all the information relevant for the recommendation process is exhaustively encoded in the knowledge base.[11] In this case it is possible to describe the profiles of both the items of the supply side to be recommended and the users at the demand site as

[9] A few hundred thousand word sample of reports, brochures, analysis etc. in green technology and company descriptions.

[10] All of the above resources were also used in the development of lexical templates discussed in Section 4.3.1

[11] Clearly this cannot be measured directly, only by evaluation of the recommended items.

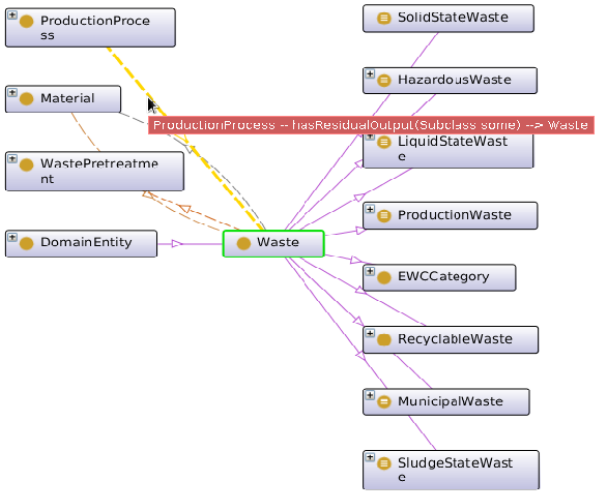

Figure 3. A fragment of the knowledge base ontology.

sets of concepts of the ontology. Basically we have on both sides a representation as a profile vector with ontology nodes as coordinates, and values as weights attached to these coordinates. In the next sections we describe how these representations are calculated.

4.2 Creating the item profiles

As we already pointed out the magnitude of the *items* (services, solutions etc.) to be recommended by the system is not so high as to exclude precise and careful manual annotation. The same is true for the rate of the inclusion of future prospective solutions into the item set. We agree here with [30] in claiming that manual annotations are more reliable than automatic ones, and when available should prevail but at the same time we leave open the possibility of running the automatic annotation process that has been developed for user data (see Section 4.3) for solutions/services but only if substantial amount of descriptive text is available about them. Otherwise it will in any case require human processing to assign the proper concept set to the items by collecting the necessary information from any sources possible.[12] The automatic annotation, if applicable, returns an initial concept set[13], to be revised and modified as needed by the human annotator. In the annotation of the items the general principle to be followed is to assign the most specific relevant concepts along *is-a* hierarchies from the knowledge base. This will constitute the default semantic profile of an item, in which weights can be set in accordance with the relevance of the given concept (node). In our initial setup, all weights are set to 1. All further related nodes will be taken care of during profile extension, which is run for the item as well as for the user profiles (see Section 4.4).

Currently we have more than one hundred different items encoded in the system which is not too many compared to other recommendation domains but can already have a significant contribution for companies in finding new economic ways to reduce their expenditures.

4.3 Creating the user profiles

User profile calculation is a conversion from available user data almost exclusively in the form of unstructured natural language Hungarian text to the vector space specified by the knowledge base. It is possible to incorporate other information sources into the user profile, which is briefly discussed in Section 4.6, but the primary task is the analysis and processing of a set of text documents for a given user. This is carried out in a standard way (see for example [26] for a similar method, although used for item annotation in their system) using Apache Tika[14] for document preprocessing and Hungarian NLP tools [31, 32] for linguistic analysis. The result of this step is a morphosyntactically annotated format, with detailed information of part-of-speech, inflection, stems and basic document and sentence structure. This serves as input to the actual conversion, for which we developed an independent lexical resource layer.

[12] It could in principle be possible to develop a separate mining module for this task but the effort invested would hardly be returned, again mostly because of the relatively low number of candidate items.

[13] As we will soon see the same process will create the profile representations for users.

[14] `https://tika.apache.org/`

4.3.1 Template based mapping to the KB

The constrained domain allows the more specific and less error prone mapping from text to ontology concepts by the help of predefined lexical templates as opposed to fully automatically deriving a semantic representation of documents [33] (called document annotation in for example [7]). These templates constitute an independent lexical layer and are not simple "annotation labels" in the ontology [30] and so have more expressive power than just keywords and their "lexical variants", providing more flexibility and better maintainability.
There are efforts on the standardisation of linking lexical resources to ontologies like the *lemon* model [34] but for the sake of faster development we opted for a simpler approach and template representation. The DIAL language, as an elaborated information extraction language allowing for complex templates for concepts whose instances are to be found in the text [35] can also be considered as a possibility to work with, however we think that very complex templates are difficult and time consuming to create and so we again voted for simplicity but scalability in this respect; if necessary more matching conditions can be added. Figure 4 illustrates a default template frame. Some of the fields may contain lists, in this case list members are interpreted conjunctively: all must be matched against the input text to activate the given concept. Disjunctive interpretation is achieved by defining several templates for one concept node.
Similarly to the process used by [36] and [34], the default set of templates are populated automatically by shallow parsing the occasionally available ontology labels or short descriptions attached to imported predefined taxonomies (like EWC and HuNACE), using the Hungarian WordNet to extend the templates from matching synsets. The generation is based on some simple heuristics referencing the linguistic annotation. Manual examination, extension and correction is necessary to ensure template reliability and high quality matching by filtering out the unnecessary noise and adding further specifications in the template fields. In the current development stage more than 5000 templates are used.

```
<#concept_ID> = [
  key = {unique template identifier}
  description = {"Descriptive text"}
  token = {(list of) wordform(s)}
  stem = {(list of) stem(s)}
  pos =  {Part-of-speech of stem(s)}
  infl = {inflectional suffix(es)}
  context = {other related stem(s)}
  context_domain =
   {domain in which to
    search for context}
  relation =
   {syntactic relation
    btw stem and context}
  weight = {1}
]
```

Figure 4. Lexical mapping template scheme used in the system.

4.3.2 Setting the vector coordinates

The template parser is run on the user documents analysed by the NLP tools and tries to match as many of the lexical templates as possible searching in the input text for the patterns specified by the template. If a match is found the KB concept as defined by the concept ID (see Figure 4, first line) of the template is activated and the corresponding coordinate of the profile vector is set to the value of the weight field of the template. This value is 1 by default but can be altered by the human annotator. Since an arbitrary number of templates (patterns) can be attached to one KB concept, it is possible that more than one template will activate one KB node, here the weights are simply added in this stage and it is the task of the profile extension (Section 4.4) to handle this case in one of the many possible ways.
Note, that contrary to the approach taken for the item annotation, templates are not only attached to the lowest leaf KB concept in an *is-a* hierarchy. There are (less specific) templates defined for more general concepts as well so that the mapping can establish links from document content to as many relevant concepts as

possible.[15]

4.4 Profile extension

The extension of the default initial KB based user and item profiles exploiting the semantic relations defined in the ontology has been shown many times to improve recommendation [14]. In our system the expansion is necessary to find a better match between the profiles by utilising the semantic context of the concepts activated in the default vectors. This context is represented as the set of entities directly linked in the ontology by explicit relations.

The primary relation that governs the hierarchy of the concepts in the ontology is the *is-a* (`SubClassOf`) relation. In our current ontology there are about 50 further relations to represent the various semantic links among concept nodes. The input to the CSA algorithm is the initial nodeset created as described in Section 4.2 and 4.3, with weights assigned to each node (default is 1.0). The output is the list of all activated nodes with weight values calculated by the algorithm.

The ontology being in OWL format, we use the Tawny-OWL toolset [37] for the necessary operations on the ontology. The built in Hermit reasoner is run to process the inferences on the base ontology and then all the members of the initial node set are queried for neighbours through all attached relations. If a concept is defined by a complex expression (equivalent class), then atomic units are extracted from this expression, which are of the format 'relation+node'. For example, the `#LiquidEnergySource` (like fuel) concept is equivalent to `#EnergySource and (#hasStateOfMatter some (#Liquid))`. From this expression two units are extracted: `#SubClassOf+#EnergySource`; `#hasStateOfMatter+#Liquid`. Based on this information we can infer that `#LiquidEnergySource` is directly related to two other nodes: its direct parent is `#EnergySource`, and it is also connected to `#Liquid` through the `#hasStateOfMatter` relation.

Weights are also assigned to each relation in the ontology, reflecting the importance of the relation and the preference for a particular path to solve a desired task. Setting these weights is a difficult knowledge engineering challenge and fine tuning is necessary through extensive testing. Currently all relations (*is-a* and all others) are assigned a default weight of 1.0.

The activation weight of a node (j) which is activated through a relation from node (i) is calculated as follows:

$$w_j = \max_i [w_i \cdot w_{ij} \cdot (1 - \alpha)] \qquad (1)$$

where w_i is the weight of the initial node, w_{ij} is the weight of the relation between i and j, and α is a decay factor. If a node is activated from multiple directions, there can be several ways of accumulating the weights of the input paths, with (interpolated) addition being a standardly used method [14,22]. Here in (1) we opt for using the maximal value but this setting is open for empirical testing. The above formula is applied recursively to newly activated nodes resulting in the extended node sequences of the profiles.

In order to control the propagation of activation out of the several possible constraints a distance threshold [22] is applied in our system.

4.5 Matching profiles and ranking the results

A straightforward consequence of the architecture of our system is that the main difficulty and focus of the recommendation task is the specification of profile representations in the common semantic, conceptual space. Once this is available, however, finding the most interesting items to be recommended is a simple cosine similarity calculation (2) between the user ($\vec{u}$) and item ($\vec{i}$) profile vectors, and the ranking of results naturally follows from the

[15] Even if we cannot calculate a mapping from a document related to for example waste treatment to the *MunicipalWaste* node, it is important to find a potential link to the more general *Waste* node.

similarity values.

$$\cos(\vec{u},\vec{i}) = \frac{\vec{u}\cdot\vec{i}}{|\vec{u}||\vec{i}|} = \frac{\sum_{i=1}^{n} u_i i_i}{\sqrt{\sum_{i=1}^{n} u_i^2} \times \sqrt{\sum_{i=1}^{n} i_i^2}} \quad (2)$$

An alternative approach would be to translate from the semantic user profiles into formal queries using languages such as RDQL or SPARQL. It can be argued that the high formalisation and expressivity available in these queries would lead to more precise results but this method would give rise to a number of new problems which are not trivial to resolve, including the translation itself.

It must be noted that in standard recommendation terminology the *degree of interest* (DOI) of the user is expressed with the concept vector weights, since obviously due to the lack of user interaction data it cannot be calculated from user behaviour. Also, context-awareness of the recommendation situation is implicitly defined and implemented through the knowledge base: anything that can be contextually relevant for the recommendation can be encoded in the ontology and if identified in the source data, will appear in the user profile vector.

4.6 Extensions to the default approach

4.6.1 Hard constraints as biases

The coordinates of the extended user profile vector can be explicitly manipulated or set by various constraints [38] which can be defined at the level of the ontology, the semantic analysis of the user data or even over the set of item representations.[16] Another source of introducing biases into the framework is the application of short audit forms which are activated by specific concept nodes in the user profile vector and present a few multiple choice questions (with preset defaults) to the user of the system asking for some more specific information. If the user is willing to give answers her profile is adjusted accordingly or some other measure is taken (for example demotion of some originally recommended item etc.) as defined for the specific answers in the specific form.

[16] In a similar vein to *business rules* in classic online market recommendation applications.

4.6.2 Incorporating relevance feedback

A nice advantage of the vector space representation of user profiles is that relevance feedback can be incorporated straightforwardly adapting a standard update formula [39] in the following way. Let's assume that $\vec{u}_i$ represents the user's original profile, for which a set of recommended items above some similarity threshold is returned. R denotes the set of items labelled relevant by the user, S is the set of non-relevant items, $\vec{r}$ and $\vec{s}$ represent the items in these two sets, respectively. There are two empirically adjusted parameters β and γ as weighting factors in the formula such that $\beta + \gamma = 1$. Now the updated user profile is calculated as in (3):

$$\vec{u}_{i+1} = \vec{u}_i + \frac{\beta}{R}\sum_{j=1}^{R}\vec{r}_j - \frac{\gamma}{S}\sum_{k=1}^{S}\vec{s}_k \quad (3)$$

5 EVALUATION

Evaluation of recommender systems is a notoriously difficult undertaking and there are many problems and unresolved issues with respect to standardisation of data and methods [40]. In the present stage of development a standard *precision at* N evaluation (4) can be performed with the top 5, 10, 15 recommended items with four different experimental scenarios: i) the baseline uses a simple keyword overlap between item and user text; ii) *[no CSA]* is default vector similarity without spreading; iii) *[CSA s-a]* is similarity with spreading activation along *is-a* relation only; iv) *[CSA all]* is spreading along all relations.

$$P@N = \frac{c(\text{relevant items in the top N items})}{N} \quad (4)$$

First results of preliminary tests are presented in Figure 5 and seem to be in line with the expectations: the more powerful models will have better performance.

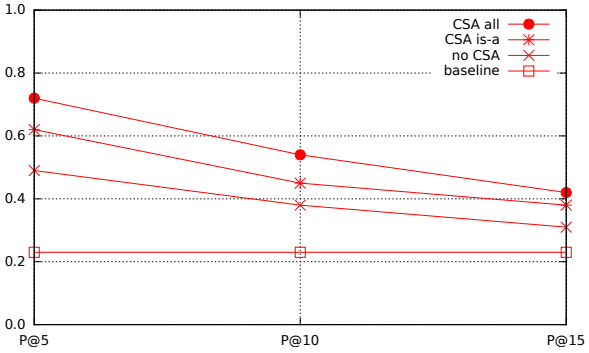

Figure 5. Evaluation results of the experiments.

6 CONCLUSIONS AND FURTHER WORK

We presented the development and use of linguistic and semantic resources to build a recommender for a novel domain in a low data density situation. The knowledge base ontology is used as an enhanced representation of the relevant knowledge about the domain of discourse, about users, about contextual conditions, while NLP techniques together with lexical resouces help in the semantic analysis of source data to provide a representation in common semantic space, in which similarity and ranking become a problem with a straightforward IR solution.

A clear advantage is the proactivity of the system: users do not need to know anything about the domain, the recommended items, only minimal user activity is required. The shallow linguistic processing of input text is quick, the system works robustly and returns a result even for minimal input information. We have also shown that NLP methods can contribute to accumulate valuable user data from available free text documents.

There remain several shortcomings and ample room for further improvements. We do not yet exploit the full potential of an ontological language, and more complex linguistic processing and so more complex relation and information extraction is also possible as is the linking of the ontology to higher level standard ontologies. The system operates in a limited domain with a limited scope and we have to live with the fact that we cannot expect usable amount of user interaction to base any recommendation strategy on, there is little if any chance of more personalization of user profiles, the minimal feedback expectable is best utilized in the form of the hard constraints intruduced into the recommendation.

ACKNOWLEDGEMENTS

This work has been supported by the National Research Development and Innovation Fund under contract KMR_12-1-2012-0036. For ontology management we use the Protege framework (`http://protege.stanford.edu/`).

REFERENCES

[1] F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, Eds., *Recommender Systems Handbook*. Springer, 2011.

[2] R. Burke, “Hybrid recommender systems: survey and experiments,” *User Modeling and User-Adapted Interaction*, vol. 12, no. 4, pp. 331–370, 2002.

[3] ——, “Knowledge-based recommender systems,” in *Encyclopedia of Library and Information Systems*, A. Kent, Ed. New York: Marcel Dekker, 2000, vol. 69.

[4] V. Codina and L. Ceccaroni, “Semantically-enhanced recommender systems,” in *Proceedings of the International Congress of the Catalan Association of Artificial Intelligence*, D. Riaño, E. Onaindia, and M. Cazorlam, Eds. Alicante, Spain: IOS Press, Amsterdam, The Netherlands, 2012, pp. 69–78.

[5] S. E. Middleton, D. De Roure, and N. R. Shadbolt, “Ontology-based recommender systems,” in *Handbook on Ontologies in Information Systems*, S. Staab and R. Studer, Eds. Springer, 2008.

[6] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. Harlow: Addison Wesley, 1999.

[7] P. Castells, M. Fernández, and D. Vallet, “An adaptation of the vector-space model for ontology-based information retrieval,” *IEEE Transactions on Knowledge and Data Engineering, Special Issue on "Knowledge and Data Engineering in the Semantic Web Era"*, vol. 19, no. 2, pp. 261–272, 2007.

[8] D. Vallet, M. Fernández, and P. Castells, "An ontology-based information retrieval model," in *Proceedings of the 2nd European Semantic Web Conference (ESWC 2005)*, Heraklion, Greece, 2005, pp. 455–470.

[9] S. E. Middleton, N. R. Shadbolt, and D. De Roure, "Ontological user profiling in recommender systems," *ACM Transactions on Information Systems*, vol. 22, no. 1, pp. 54–58, 2004.

[10] A. Al-Nazer, T. Helmy, and M. Al-Mulhem, "User's profile ontology-based semantic framework for personalized food and nutrition recommendation," *Procedia Computer Science*, vol. 32, pp. 101–118, 2014.

[11] Y. Blanco-Fernández, J. J. Pazos-Arias, A. Gil-Solla, M. Ramos-Cabrer, M. López-Nores, J. García-Duque, A. Fernández-Vilas, R. P. Díaz-Redondo, and J. Bermejo-Muñoz, "A flexible semantic inference methodology to reason about user preferences in knowledge-based recommender systems," *Knowedged-Based Systems*, vol. 21, no. 4, pp. 305–320, May 2008.

[12] A. Sieg, B. Mobasher, and R. Burke, "Improving the effectiveness of collaborative recommendation with ontology-based user profiles," in *Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems*, ser. HetRec '10. New York, NY, USA: ACM, 2010, pp. 39–46. [Online]. Available: http://doi.acm.org/10.1145/1869446.1869452

[13] A. Hawalah and M. Fasli, "Using user personalized ontological profile to infer semantic knowledge for personalized recommendation." in *EC-Web*, ser. Lecture Notes in Business Information Processing, C. Huemer and T. Setzer, Eds., vol. 85. Springer, 2011, pp. 282–295.

[14] I. Cantador, P. Castells, and A. Bellogín, "An enhanced semantic layer for hybrid recommender systems: Application to news recommendation," *International Journal On Semantic Web and Information Systems*, vol. 7, no. 1, pp. 44–77, 2011.

[15] A. Kiryakov, B. Popov, I. Terziev, D. Manov, and D. Ognyanoff, "Semantic annotation, indexing, and retrieval," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 2, no. 1, 2004.

[16] A. Wessman, S. W. Liddle, and D. W. Embley, "A generalized framework for an ontology-based data-extraction system," in *Proceedings of the 4th International Conference on Information Systems Technology and its Applications*, 2005, pp. 239–253.

[17] N. Kiyavitskaya, N. Zeni, J. R. Cordy, L. Mich, and J. Mylopoulos, "Cerno: light-weight tool support for semantic annotation of textual documents," *Data and Knowledge Engineering*, vol. 68, no. 12, pp. 1470–1492, 2009.

[18] S. Kara, Özgür Alan, O. Sabuncu, S. Akpınar, N. K. Cicekli, and F. N. Alpaslan, "An ontology-based retrieval system using semantic indexing," *Information Systems*, vol. 37, no. 4, pp. 294 – 305, 2012, semantic Web Data Management.

[19] V. Codina and L. Ceccaroni, "Taking advantage of semantics in recommendation systems," in *Proceedings of the International Congress of the Catalan Association of Artificial Intelligence*. L'Espluga de Francolí, Spain: IOS Press, Amsterdam, The Netherlands, 2010, pp. 163–172.

[20] F. Crestani, "Application of spreading activation techniques in information retrieval," *Artificial Intelligence Review*, vol. 11, no. 6, pp. 453–482, 1997.

[21] Y. Blanco-Fernández, M. López-Nores, and J. J. Pazos-Arias, "Adapting spreading activation techniques towards a new approach to content-based recommender systems," in *Intelligent Interactive Multimedia Systems and Services*, ser. Smart Innovation, Systems and Technologies, G. A. Tsihrintzis, E. Damiani, M. Virvou, R. J. Howlett, and L. C. Jain, Eds. Springer Berlin Heidelberg, 2010, vol. 6, pp. 1–11.

[22] C. Rocha, D. Schwabe, and M. P. Aragao, "A hybrid approach for searching in the semantic web," in *Proceedings of the 13th International Conference on World Wide Web*, ser. WWW '04. New York, NY, USA: ACM, 2004, pp. 374–383. [Online]. Available: http://doi.acm.org/10.1145/988672.988723

[23] T.-P. Liang, Y.-F. Yang, D.-N. Chen, and Y.-C. Ku, "A semantic-expansion approach to personalized knowledge recommendation," *Decision Support Systems*, vol. 45, no. 3, pp. 401–412, Jun. 2008.

[24] I. Cantador, A. Bellogín, and P. Castells, "A multilayer ontology-based hybrid recommendation model," *AI Communications*, vol. 21, no. 2-3, pp. 203–210, Apr. 2008.

[25] I. Cantador, M. Fernández, D. Vallet, P. Castells, and M. R. Jérôme Picault, "A multi-purpose

ontology-based approach for personalised content filtering and retrieval," in *Advances in Semantic Media Adaptation and Personalization*, ser. Studies in Computational Intelligence, M. Wallace, M. Angelides, and P. Mylonas, Eds. Springer-Verlag, 2008, vol. 93, pp. 25–51.

[26] M. Fernández, I. Cantador, V. López, D. Vallet, P. Castells, and E. Motta, "Semantically enhanced information retrieval: An ontology-based approach," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 9, no. 4, pp. 434–452, 2011, {JWS} special issue on Semantic Search.

[27] P. Rayson and R. Garside, "Comparing corpora using frequency profiling," in *Proceedings of the Workshop on Comparing Corpora*. Association for Computational Linguistics, 2000, pp. 1–6.

[28] P. Pecina, "A machine learning approach to multiword expression extraction," in *Proceedings of the LREC2008 workshop: Towards a Shared Task for Multiword Expressions*, Marrakech, Morocco, 2008, pp. 54–57.

[29] M. Miháltz, C. Hatvani, J. Kuti, G. Szarvas, J. Csirik, G. Prószéky, and T. Váradi, "Methods and results of the hungarian wordnet project," in *Proceedings of the Fourth Global WordNet Conference*, 2008, pp. 310–320.

[30] D. Vallet, M. Fernández, and P. Castells, "An ontology-based information retrieval model," in *Proceedings of the Second European Conference on The Semantic Web: Research and Applications*, ser. ESWC'05. Berlin, Heidelberg: Springer-Verlag, 2005, pp. 455–470.

[31] Cs. Oravecz and P. Dienes, "Efficient stochastic part-of-speech tagging for Hungarian," in *LREC-02*, Las Palmas, Canary Islands, Spain, 2002, pp. 710–717.

[32] J. Zsibrita, V. Vincze, and R. Farkas, "magyarlanc: A toolkit for morphological and dependency parsing of Hungarian," in *Proceedings of RANLP*, 2013, pp. 763–771.

[33] M. Baziz, "Towards a semantic representation of documents by ontology-document mapping," in *Artificial Intelligence: Methodology, Systems, and Applications*, ser. Lecture Notes in Computer Science, C. Bussler and D. Fensel, Eds. Springer Berlin Heidelberg, 2004, vol. 3192, pp. 33–43.

[34] J. McCrae, D. Spohr, and P. Cimiano, "Linking lexical resources and ontologies on the semantic web with lemon," in *Proceedings of the 8th Extended Semantic Web Conference (ESWC)*, 2011, pp. 245–259.

[35] R. Feldman and J. Sanger, *Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. New York, NY, USA: Cambridge University Press, 2006.

[36] J. McCrae, E. Montiel-Ponsoda, and P. Cimiano, "Collaborative semantic editing of linked data lexica," in *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, 2012, pp. 2619–2625.

[37] P. Lord, "The semantic web takes wing: Programming ontologies with Tawny-OWL," in *Proceedings of the 10th International Workshop on OWL: Experiences and Directions (OWLED 2013), vol. 1080 CEUR-WS.org: CEUR Workshop Proceedings*, M. Rodriguez-Muro, S. Jupp, and K. Srinivas, Eds., Montpellier, France, 2013.

[38] A. Felfernig and R. Burke, "Constraint-based recommender systems: technologies and research issues," in *Proceedings of the 10th international conference on Electronic commerce*, ser. ICEC '08. New York, NY, USA: ACM, 2008, pp. 3:1–3:10. [Online]. Available: http://doi.acm.org/10.1145/1409540.1409544

[39] J. J. Rocchio, "Relevance feedback in information retrieval," in *The SMART Retrieval System – Experiments in Automatic Document Processing*, G. Salton, Ed. Englewood Cliffs, NJ: Prentice Hall, 1971, pp. 313–323.

[40] A. Gunawardana and G. Shani, "A survey of accuracy evaluation metrics of recommendation tasks," *The Journal of Machine Learning Research*, vol. 10, pp. 2935–2962, 2009.