Injection of Knowledge in a Sourcing Recommender System

Molka Tounsi Dhouib, Catherine Faron, Andrea G. B. Tettamanzi

Injection of Knowledge in a Sourcing Recommender System

Molka Tounsi Dhouib
Université Côte d'Azur, Inria, CNRS, I3S, Silex
Sophia Antipolis, France
0000-0002-6670-7821

Catherine Faron
Université Côte d'Azur, Inria, CNRS, I3S
Sophia Antipolis, France
0000-0001-5959-5561

Andrea G. B. Tettamanzi
Université Côte d'Azur, Inria, CNRS, I3S
Sophia Antipolis, France
0000-0002-8877-4654

***Abstract*—Recommender systems provide suggestions to users for items that best meet their needs. In this work, we study the benefits of using knowledge and, more specifically, a 'bag of concepts' representation to enhance a recommender system in the sourcing domain. We tested our approach in a real-world case study provided by the *Silex* company. The experimental results show that injecting knowledge in the recommendation process outperforms word embedding approaches.**

I. Introduction

Nowadays, Recommender systems (RS) are primarily used in commercial applications to provide a personalized experience and suggest relevant items to users such as music, movies, books, trips, products. In the context of the *Silex* company, our goal is to propose a RS that can predict relevant providers (items) that are likely to be of interest for a service request (user). *Silex* simplifies the sourcing process by allowing companies to provide a description of their professional activities, their offers and/or the services they are looking for in natural language, in French.

In this paper, we propose to combine a conceptual representation of texts to their representation based on word embedding to enhance the recommendation in the sourcing domain. Our main research question is: Can the integration of domain knowledge enhance the performance of a recommender system in our use case? We focus on the following sub-questions: (i) What is the best way to integrate domain knowledge into the representation of service requests and providers in order to enhance the quality of recommendations? (ii) To what extent does the injection of domain knowledge improve the performance of the system?

The remainder of this paper is organized as follows. Section II gives an overview of state of the art recommendation approaches. Section III presents the recommendation approach we propose. Section IV reports and discusses the results of our experiments in the sourcing domain. Section V concludes and provides an outline of future work.

II. Related Work

There are four main approaches to RS:

- Content-based recommendation [1]: This approach consists in analyzing the content (i.e., set of attributes or metadata) of items liked by users in the past and suggests items with similar content [2].
- Collaborative filtering [3]: It is based on the feedback (rating) on items provided by similar users.
- Knowledge-based recommendation [4]: This approach consists in using ontologies to model knowledge about the user context, the item context and the domain.
- Hybrid recommendation [5]: This approach combines the three above ones, to take advantage of the benefits of each of them.

Our method is at the intersection of the content-based and knowledge-based approaches. The idea behind our approach is to take advantage of the power of word embedding, which approximates a general semantic relation, and the power of domain knowledge, which models a more specific semantic relation, to provide high quality recommendations.

III. Proposed Approach

Our proposal relies on the use of domain knowledge that can be captured into a thesaurus or an ontology. In the following we will refer to vocabulary indistinctly as a thesaurus or an ontology. In the experiments reported in this paper and conducted to answer the Silex use case, we are using a sourcing domain vocabulary, as we exploit different types of relationship between concepts and their labels (and not intentional definitions).

Our approach comprises five steps: (i) construction of a vocabulary for the sourcing domain, (ii) entity recognition from the textual descriptions of service requests and providers, (iii) entity management, (iv) vector representation of service requests and providers, and (v) recommendation algorithm.

A. Vocabulary for the Sourcing Domain

The preliminary step of our approach is the construction of a vocabulary to capture the sourcing knowledge [6] by identifying and combining several relevant standard metadata repositories such as ESCO,[1] ROME,[2] NAF,[3] UNSPSC,[4] and

[1]https://ec.europa.eu/esco/portal/home
[2]http://www.pole-emploi.org/accueil/mot-cle.html?tagId=94b2eaf6-d7bd-4244-bddc-01415605563b
[3]https://www.insee.fr/fr/information/2406147
[4]https://www.unspsc.org/

CPF,[5]. Our sourcing vocabulary gathers 124040 concepts. We proposed an alignment method and applied it to automatically align the above-mentioned vocabularies, which where partially overlapping [7].

B. Sourcing Named Entity Recognizer

The second step of our approach consists in extracting the relevant entities (i.e., services, products, occupations, skills, and business sectors) from the textual descriptions (i.e. service requests and company descriptions), based on our sourcing vocabulary. To handle this type of textual data, our approach [8] is based on a Bidirectional Long-term and Short-term Mermory (Bi-LSTM) encodes an input textual description and a Conditional Random Fields model (CRF) labels every word to detect entities. In addition to word embedding, we extract three other kinds of embedding for each word in the textual description: (i) a character-level embedding to represent rare words or words with spelling errors; (ii) a syntax embedding to locate the user's need in a description: and (iii) a position embedding to push the model to understand that it is highly likely that the words at the beginning of the text are relevant information.

C. Entity Linking with the Sourcing Vocabulary

In order to link the named entities extracted from the descriptions of service requests and providers with the closest concepts in the sourcing vocabulary, we defined a similarity measure between an entity and a concept.

We first represent each extracted entity and each concept of the sourcing vocabulary by an embedding vector which is computed as the average of the word embedding vectors of all the words participating in the entity or label of concept and occurring in the dictionary. In the absence of a large corpus available to train a word embedding model for our use case, we use fastText pre-trained word vectors for French.

The embedding vector for an entity or a (label of) concept x is thus computed as

$$V(x) = \frac{1}{n}\sum_{i=1}^{n} w_i, \tag{1}$$

where n is the number of words of the dictionary occurring in x and $w_i \in \mathbb{R}^{300}$ denotes the word embedding vector of the ith word of x occurring in the dictionary. If a word of x does not belong to the dictionary, it is just ignored.

Then we define the similarity between an entity e extracted form a request or provider description and a concept c in the sourcing vocabulary as the cosine similarity between their embedding vectors $V(e)$ and $V(c)$:

$$sim(e,c) = \frac{V(e)\cdot V(c)}{||V(e)||\cdot||V(c||}. \tag{2}$$

Finally, we link each entity with the most similar concept in the vocabulary O:

$$linked_0(e,c) \iff sim(e,c) = \max_{c_i \in O} sim(e,c_i). \tag{3}$$

[5]https://www.insee.fr/fr/metadonnees/cpfr21/section/A?champRecherche=true

D. Vector Representation of Service Requests and Providers

We aim to represent each service request or provider by a vector that summarizes the semantics of the entities extracted from its description. For each description of a service request or provider, we consider three alternative vector representations: (i) the average of the embedding vectors of the entities in the textual description, (ii) a bag of concepts representation, and (iii) a vector representation combining the two former ones.

1) Word Embedding of Entities: The base vector representation $V_{Emb}(x)$ of a service request or provider x is the average of the embedding vectors of all the entities $e_i, i = 1, ..., n$ extracted from its description:

$$V_{Emb}(x) = \frac{1}{n}\sum_{i=1}^{n} V(e_i), \tag{4}$$

where $V(e_i)$ is the vector representation of entity e_i as defined in "(1)".

2) Bag of Concepts: Using the result of the above described entity linking process, we consider an alternative representation $V_{BoC}(x)$ of a service request or provider x based on the sourcing vocabulary S: the bag of the concepts (BoC) in S which the entities e_i extracted from x are linked to according to "(3)":

$$V_{BoC}(x) = BoC_S(x) = (b_1, ..., b_m) \tag{5}$$

where m is the size of the sourcing vocabulary S and $b_i = 1$ if $\exists e \in x, linked_S(e, c_i)$, and $b_i = 0$ otherwise. Additionally, we considered enriching the BoC representation of a service request or provider, by considering not only the concepts linked to the entities it contains but also some neighbors in the vocabulary to the linked concepts. More precisely, we considered the parents of the concepts linked to the entities (skos:narrower relation) or those semantically close (skos:closeMatch relation). Formally, we define three alternative BoC representations:

$$V'_{BoC}(x) = BoC'_S(x) = (b_1, ..., b_m) \tag{6}$$

where m is the size of the sourcing vocabulary S and $b_i = 1$ if $\exists e \in x, linked_S(e, c_i)$ or $linked_S(e, c_j)$ with c_j skos:narrower c_i; and $b_i = 0$ otherwise.

$$V''_{BoC}(x) = BoC''_S(x) = (b_1, ..., b_m) \tag{7}$$

where m is the size of the sourcing vocabulary S and $b_i = 1$ if $\exists e \in x, linked_S(e, c_i)$ or $linked_S(e, c_j)$ with c_j skos:closeMatch c_i; and $b_i = 0$ otherwise.

$$V'''_{BoC}(x) = BoC'''_S(x) = (b_1, ..., b_m) \tag{8}$$

where m is the size of the sourcing vocabulary S and $b_i = 1$ if $\exists e \in x, linked_S(e, c_i)$ or $linked_S(e, c_j)$ with c_j skos:narrower c_i or c_j skos:closeMatch c_i; and $b_i = 0$ otherwise.

3) Combination of Vector Representations: We define a third type of vector representation of a description of service request or provider as the concatenation of the vector representations defined in "(4)" and one of the BoC representations defined in "(5)", "(6)", "(7)", and "(8)", respectively:

$$V_{Conc}(x) = V_{Emb}(x)^\frown V_{BoC}(x), \quad (9)$$
$$V'_{Conc}(x) = V_{Emb}(x)^\frown V'_{BoC}(x), \quad (10)$$
$$V''_{Conc}(x) = V_{Emb}(x)^\frown V''_{BoC}(x), \quad (11)$$
$$V'''_{Conc}(x) = V_{Emb}(x)^\frown V'''_{BoC}(x). \quad (12)$$

E. Recommendation Algorithm

We define two metrics to measure the similarity between a service provider p and a service request r. The first one is the cosine similarity between the vector representations of p and r:

$$sim_1(p,r) = \frac{V(p) \cdot V(r)}{||V(p)|| \cdot ||V(r||}, \quad (13)$$

where V stands for one of the representations defined in "(4)" to "(12)".

The second metrics is

$$sim_2(p,r) = \begin{cases} 1, & \text{if } V(p) \cap V(r) \neq 0, \\ sim_1(p,r), & \text{otherwise,} \end{cases} \quad (14)$$

where sim_1 is then computed with the base vector representation V_1.

A service provider p is recommended for a service request r if $sim_1(p,r)$ or $sim_2(p,r)$ is greater than a given threshold, depending on the chosen similarity measure.

IV. Experiments and Results

A. Dataset and Protocol

We evaluate the performance of our recommendation approach on a dataset provided by sales experts at Silex. This dataset comprises 109 descriptions of service requests and the 649 providers in various areas which were manually selected and recommended for these requests. We consider two sets of annotations of this dataset: A is the set of annotations automatically performed by our NER approach; A' is the result of a manual cleaning of A that we performed.

To decide on the optimal vector representation and algorithm to recommend service providers, we conducted eleven experiments whose settings are depicted in Table I, the baseline being experiment Emb.

TABLE I: Experimental settings

Experiment	Vector representation	Similarity measure
Emb	V_{Emb}	sim_1
BoC	V_{BoC}	sim_1
BoC′	V'_{BoC}	sim_1
BoC″	V''_{BoC}	sim_1
BoC‴	V'''_{BoC}	sim_1
Conc	V_{Conc}	sim_1
Conc′	V'_{Conc}	sim_1
Conc″	V''_{Conc}	sim_1
Conc‴	V'''_{Conc}	sim_1
BoC→Emb	V_{BoC} and V_{Emb}	sim_2
Conc→Emb	V_{Conc} and V_{Emb}	sim_2

In order to evaluate the performance of the proposed settings, and therefore the interest of injecting domain knowledge into vector representations, we used the precision, recall, F1 score metrics and the precision score considering the N top ranking providers (up to the tenth) according to the usual formula for "precision at N": .

$$P@N = \frac{\text{relevant items in the top } N \text{ recommended items}}{N}. \quad (15)$$

We conducted the parameter learning (i.e threshold) through 5-fold cross validation.

B. Results and Discussion

Table II and III present the performance of our system for each tested setting in terms of precision, recall and F1 score with datasets A and A' respectively.

TABLE II: Evaluation of the proposed experimental settings with dataset A.

Experiment	Threshold	Precision	Recall	F1
Emb	0.76	0.848	0.397	0.530
BoC	0.239	**0.887**	0.137	0.235
BoC′	0.16	0.852	0.247	0.378
BoC″	0.15	0.728	0.187	0.296
BoC‴	0.069	0.697	0.287	0.402
Conc	0.58	0.875	**0.612**	**0.717**
Conc′	0.55	0.793	0.457	0.575
Conc ″	0.58	0.791	0.231	0.365
Conc ‴	0.52	0.797	0.258	0.389
BoC→Emb	0.70	0.600	0.580	0.579
Conc→Emb	0.24;0.76	0.83	0.429	0.558

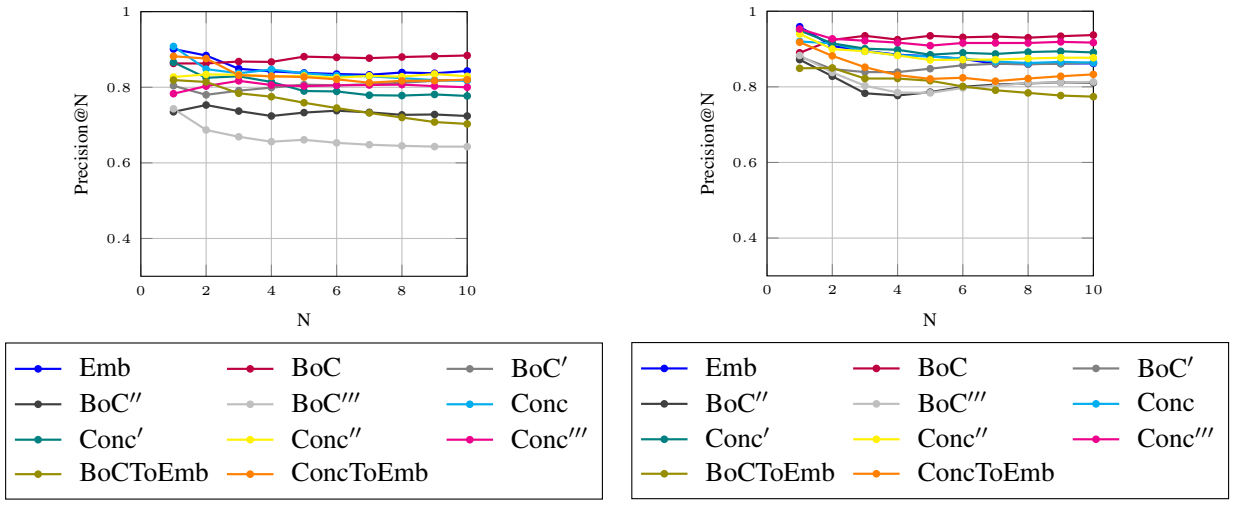
TABLE III: Evaluation of the proposed experimental settings with dataset A'.

Experiment	Threshold	Precision	Recall	F1
Emb	0.73	0.877	0.562	0.678
BoC	0.01	0.909	0.338	0.487
BoC′	0.16	0.890	0.456	0.596
BoC″	0.13	0.819	0.414	0.541
BoC‴	0.10	0.841	0.497	0.618
Conc	0.58	0.875	0.612	0.717
Conc ′	0.55	0.912	0.540	0.674
Conc ″	0.55	0.883	0.420	0.568
Conc ‴	0.52	**0.920**	0.430	0.584
BoC→Emb	0.7	0.72	**0.694**	**0.768**
Conc→Emb	0.05;0.73	0.860	0.627	0.719

In order to evaluate to what extent the performance of our proposed approach depends on the method adopted to automatically annotate the descriptions of the service requests and providers, we conducted some additional experiments in which we used DBpedia spotlight[6] for NER. Table IV presents the results on dataset A when named entities are extracted with DBpedia spotlight [9], [10].

Figures 1a and 1b present the performance of our system for each tested setting in terms of P@N on dataset A and dataset A' respectively.

[6]https://www.dbpedia-spotlight.org/



(a) Dataset A.

(b) Dataset A'.

Fig. 1: Precision@N

TABLE IV: Evaluation of the proposed experimental settings with dataset A, using DBpedia spotlight for NER.

Experiment	Threshold	Precision	Recall	F1
Emb	0.79	0.346	0.169	0.219
BoC	0.11	**0.418**	**0.313**	**0.357**
Conc	0.39	0.401	0.193	0.260

All methods, where a word embedding representation is enriched with domain knowledge, obtain a better precision measure than the baseline method using word embedding alone. Focusing on the precision@N results, with both datasets, we can conclude that injecting domain knowledge is highly beneficial to our RS. Although a BoC representation performs very well from the 2 top-ranking items on dataset A, we can observe that its performance keeps steady on both datasets up to the tenth item. Using this cleaned dataset, the performance of all methods are increased. This emphasizes the fact that all methods are very sensitive to the quality of entity linking.

All in all, it appears that injecting domain knowledge into the vector representations is all the more beneficial the greater the quality of the annotations. Also, enriching the conceptual representations by considering the subsumption relation and skos:closeMatch relation clearly gives a better results. Finally, the comparison with the experiments using DBpedia spotlight for NER confirms that the introduction of domain knowledge in the recommendation process is beneficial and helps enhance the performance of the system even when one cannot rely on a high-quality NER method.

V. Conclusion

In this paper, we presented a sourcing recommender system based on domain knowledge. We focus especially on the vector representation of the descriptions by evaluating the performance of the system using word embedding or injecting domain knowledge into the representation. Our experiments show that recommendation accuracy can be greatly improved through the injection of domain knowledge in the recommendation process. As future work, we aim firstly to use BERT or Camembert for French texts as models to generate word embedding representations.

References

[1] K. Lang, "Newsweeder: Learning to filter netnews," in *Machine Learning Proceedings 1995*. Elsevier, 1995, pp. 331–339.

[2] P. S. Yu, "Data mining and personalization technologies," in *Proceedings. 6th International Conference on Advanced Systems for Advanced Applications*. IEEE, 1999, pp. 6–13.

[3] U. Shardanand and P. Maes, "Social information filtering: algorithms for automating "word of mouth"," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 1995, pp. 210–217.

[4] A. Ameen, "Knowledge based recommendation system in semantic web -a survey," *International Journal of Computer Applications Volume 182 –No. 43*, 2019.

[5] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE transactions on knowledge and data engineering*, vol. 17, no. 6, pp. 734–749, 2005.

[6] M. Dhouib, C. F. Zucker, and A. Tettamanzi, "Construction d'ontologie pour le domaine du sourcing," in *29es Journées Francophones d'Ingénierie des Connaissances, IC 2018*, 2018, pp. 137–144.

[7] M. T. Dhouib, C. F. Zucker, and A. G. Tettamanzi, "An ontology alignment approach combining word embedding and the radius measure," in *International Conference on Semantic Systems*. Springer, 2019, pp. 191–197.

[8] H. Daoud, M. T. Dhouib, J. Rancati, C. F. Zucker, and A. G. Tettamanzi, "A hybrid bi-lstm-crf model for sequence labeling applied tothe sourcing domain," *6ème conférence sur les Applications Pratiques de l'Intelligence Artificielle APIA2020*, 2020.

[9] P. N. Mendes, M. Jakob, A. Garcia-Silva, and C. Bizer, "Dbpedia spotlight: Shedding light on the web of documents," in *Proceedings of the 7th International Conference on Semantic Systems (I-Semantics)*, 2011.

[10] J. Daiber, M. Jakob, C. Hokamp, and P. N. Mendes, "Improving efficiency and accuracy in multilingual entity extraction," in *Proceedings of the 9th International Conference on Semantic Systems (I-Semantics)*, 2013.