Information Services & Use 34 (2014) 325–330
DOI 10.3233/ISU-140755
IOS Press

Similarity between text and RDF

Marcelo Schiessl [a,*], Rita Berardi [b] and Marisa Bräscher [c]

[a] *Universidade de Brasilia, Brasília, DF, Brazil*

[b] *Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, RJ, Brazil*

[c] *Universidade Federal de Santa Catarina, Florianópolis, SC, Brazil*

Abstract. Recently, sources of structured and unstructured data have been made available on the web, and gained attention among researchers from several areas. They are become more interested in using this global dataset due to its size and variety of information. In the Semantic Web field, many studies have translated structured data into unstructured data, and vice-versa, to make them comprehensible to machines and humans. However, we argue that we can take advantage of the existing information, in both text and RDF format. In this paper we focus on finding a way to compare them, and discovering which available text can represent an existing RDF. Hence, we propose a strategy to check whether a text represents the same knowledge that is shown in RDF format.

Keywords: Semantic Web, natural language processing, similarity measurement, RDF, knowledge representation

1. Introduction

Information is everywhere in a variety of formats. The Linked Data initiative has brought new opportunities for exploring data in terms of speed and efficiency. The Linked Open Data cloud (LOD) provides great sources of structured data in RDF (Resource Description Framework) format that can be consumed by humans and machines. Although this new technology is important, text is still one of the main ways of communicating and storing information.

On the one hand, the LOD cloud aggregates efficiency and speed to the access of information. On the other hand, texts are rich sources of information since they are represented in natural language that is more expressive than machine-readable languages, such as RDF. Ideally, if we had sources of information in RDF and textual formats, we could take advantage of the benefits of each format. As evidence of the importance of having an RDF version of texts, we have Wikipedia as a global encyclopedia in text, which has a part translated to RDF to be accessed by machines and can be found at DBpedia. This correspondence between Wikipedia and DBpedia is a successful example of how to represent the same knowledge in different formats, which can be consumed for different tasks [12]. Wikipedia articles are composed mostly of free text, and many efforts are made in order to extract information from there and store them in machine-readable triplestores.

Unfortunately, the correspondence between these two formats is not always straightforward. Thus, the task of translating the complex natural language (text) into a simple structure machine-readable (RDF) must be improved. The works [9] and [5] have influenced researchers in finding ways to translate natural language into machine-readable information. They use linguist patterns and statistical tools to automatically detect potential machine-readable structures. From RDF into text, several techniques are

[*] Corresponding author: Marcelo Schiessl, Universidade de Brasilia, DF, Brazil. E-mail: schiessl@unb.br.

0167-5265/14/$27.50 © 2014 – IOS Press and the authors.

being adapted or created to extract linguistic entities from structured data and transform them into a human-readable format such as in [7,11].

Although this conversion is important, why not explore the sources that already exist in both formats and try to find similarities between the information represented by these different formats? We argue that we can take advantage of the existing information, in both text and RDF formats, finding a way to compare them, and discovering which available text can represent an existing RDF. Hence, we propose a strategy to check whether a text shows the same information that is encoded in RDF format. This comparison accelerates the process of having the same information in text and RDF since it finds in advance a text to represent the RDF and both may be enriched or updated with new information. In addition, this comparison helps to decipher how patterns in natural language are represented in RDF format and vice-versa, contributing to an improvement in conversion techniques.

The remainder of this paper is organized as follows. In Section 2 we present the general background required to frame our approach. We introduce the similarity measurement problem in Section 3. Then, we present our approach in Section 4. In Section 5, we outline the experiments and discuss its results. In Section 6, the related work. Lastly, the Conclusion and future works.

2. Preliminaries – Text and RDF

We assume the reader's familiarity with RDF[1] structure and linguistic similarities techniques. In the vision of the Semantic Web, described by [3], RDF is the building block of it. In order to represent open data, W3C recommends the Linked Data standard [4], which lays out data in the form of a set of RDF triples. W3C defines RDF as a language for representing information in the Web. This language is composed by a vocabulary described in the RDF schema.[2] Any expression in RDF is a triple with a subject, a predicate and an object, each one represented by a vocabulary term. The object may be a Uniform Resource Identifier (URI), a literal or a blank node. When comparing texts and RDF data we refer only to RDF data with literal objects, because it is the part of the RDF sentence that is most similar to natural language. Here we refer to the literal RDF as structured data.

Texts are represented by natural language that is governed by a grammatical structure. Throughout this paper we consider literal RDF as structured data and unstructured all information that is not machine-readable. According to this point of view, text is considered unstructured and RDF is structured.

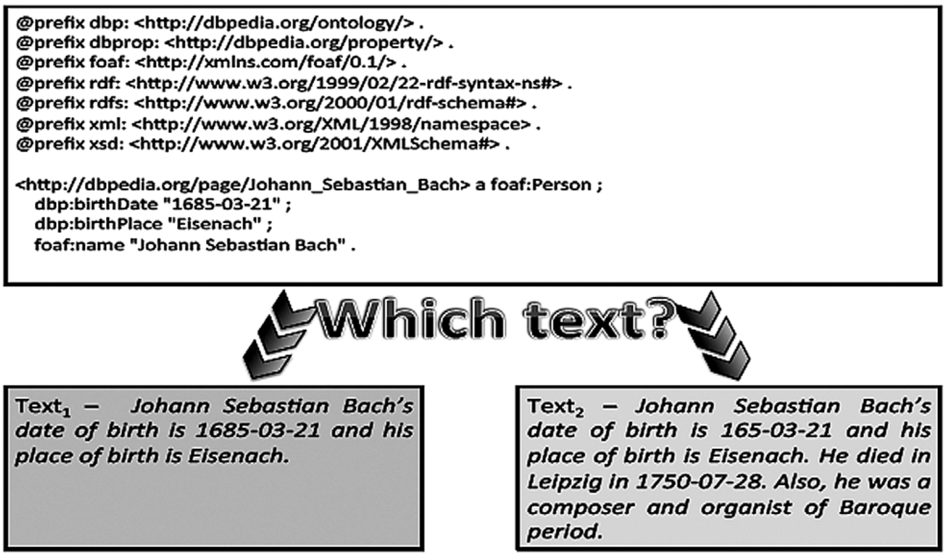
To sum up, textual and RDF data intend to represent some domain and to provide information to humans and computers, but they have different structures. Each structure has its own development and the creation of contents using them might be different. Consequently, the relationship between these two formats might be not a trivial task to discover.

3. The similarity measure problem

In this paper we focus on one direction of the similarity, summarizing the measurement problem in the following question "Can you define which text better represents the information contained in an RDF?" To explore the problem, consider the motivating example in the Fig. 1. Text 1 can be considered as having a "total" similarity since every term of the RDF structure can be found in the text in natural language.

[1] http://www.w3.org/TR/rdf-concepts/.

[2] http://www.w3.org/TR/rdf-schema/.



```
@prefix dbp: <http://dbpedia.org/ontology/> .
@prefix dbprop: <http://dbpedia.org/property/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xml: <http://www.w3.org/XML/1998/namespace> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

<http://dbpedia.org/page/Johann_Sebastian_Bach> a foaf:Person ;
   dbp:birthDate "1685-03-21" ;
   dbp:birthPlace "Eisenach" ;
   foaf:name "Johann Sebastian Bach" .
```

Text$_1$ – *Johann Sebastian Bach's date of birth is 1685-03-21 and his place of birth is Eisenach.*

Text$_2$ – *Johann Sebastian Bach's date of birth is 165-03-21 and his place of birth is Eisenach. He died in Leipzig in 1750-07-28. Also, he was a composer and organist of Baroque period.*

Fig. 1. RDF and texts.

For instance, the following terms from Text 1 "*Johann Sebastian Bach", "birth", "date", "place"* can be found in the RDF, what this means is that this structured information (RDF) is also represented by the unstructured information (Text 1). However, despite having the correspondence between the RDF data terms and the text, in Text 2 there is more information than is represented in the RDF, such as the *city* and *date* where Johann Sebastian Bach *died*. Thus, the text would be completely represented on the structured format only if more RDF data were aggregated. Based on this example, how can we objectively decide which text represents the RDF data more accurately?

4. Approach for estimating the similarity

We outline the approach for comparing RDF data with texts in order to identify which text corresponds to an RDF data. The task is to check whether the RDF triples can be somehow found in the text.

Our approach performs the similarity estimation task by considering only the triples where the object is a *literal type*, and only this part of a dataset is compared to a text. Then, the similarity can be calculated by taking a bag-of-words from both data, and applying the well-known text similarity measurements, like Dice Coefficient or Cosine Similarity [10], to quantify how similar they are. To exemplify the approach, we calculate the similarity of the example in Fig. 1 and show the results in the Table 1.

5. Experiments

For the experiment, we focused on music composers' biographies domain. The experiment consists of setting up a corpus with 3 different textual documents that correspond to an RDF file and detecting which text is more related to this RDF. Regarding the RDF, we extracted them from DBpedia. The 3 textual

Table 1

Similarity approach for the motivating example

Text × RDF	RDF literals “Johann_Sebastian_Bach”, “1685-03-21”, “Eisenach”	
	Cosine similarity	Dice coefficient
Text 1: “Johann_Sebastian_Bach’s birth of date is 1685-03-21 and his place of birth is Eisenach”	0.707	0.600
Text 2: “Johann_Sebastian_Bach’s birth of date is 1685-03-21 and his place of birth is Eisenach. He died in Leipzig in 1750-07-28. Also, he was a composer and organist of Baroque period”	0.463	0.333

documents, we extracted from Wikipedia[3] – simple English version – and from the sites[4] Naxos and Encyclopedia Britannica. We collected 7 arbitrary names of music composers, which are from different eras and countries. They are Heitor Villa-Lobos (HV), John Cage (JC), Johann Sebastian Bach (JSB), Claude Debussy (CD), Ludwig van Beethoven (LvB), Richard Wagner (RW) and Wolfgang Amadeus Mozart (WAM).

Concerning the preprocessing task, the following approaches were used in the experiments: (i) tokenization: splitting strings into substrings based on delimiters, (ii) split compounds words, e.g., “mother-in-law” or “European Union”, (iii) normalization: making words canonically equivalent, by removing differences due to capitalizations, punctuation, word representation, e.g., “Jan, 21st 2013” to “2013-01-21”, (iv) stemming/lemmatization: reducing inflectional and derivationally related forms of a word to a common word, e.g. verb tense, plurals etc., (v) stop words removal: removal of frequently used words, and (vi) part-of-speech: tagging the category of words which denotes their functions in sentences such as subjects, objects, verbs etc.

During the preprocessing, we tested four stemming methods. The highest similarity was reached by using Lancaster stemmer. We used cosine similarity, which varies from 0 – completely different – to 1 – the exact copy – to compare the documents. Every date was converted to the format “yyyy-mm-dd”. The part-of-speech was used to disambiguate words. Named Entity were converted to the form firstName_secondName, e.g., Johann Sebastian Bach to Johann_Sebastian_Bach. Each RDF was compared to all texts from the corpus. The total of comparisons was 147 times, which corresponds to 1 RDF compared to the 3 versions of each txt file, i.e., Wikipedia, Naxos and Britannica. After the preprocessing, the texts collection is composed of 28 files, 191,704 characters, 31,267 words, 1690 sentences and 564 distinct words of which the vocabulary is comprised.

5.1. Results

Figure 2 shows the result for the experiments. We present only the four highest similarity measures.

Each group represents the RDF and the four bars, the texts corresponding to the highest similarity. As an illustration of how to interpret the graph, the first group is related to HV (Heitor Villa-Lobos), and the bars indicate degree of similarity.

For every group, the highest similarity measurements corresponded to the same composer in the RDF. In other words, the RDF of *Johann Sebastian Bach* has the highest similarity with a text related to *Johann Sebastian Bach*, and so on. Most of the top measures are above 0.6, which indicates an acceptable

[3] http://simple.wikipedia.org/wiki/Main_Page.

[4] http://www.naxos.com, http://www.britannica.com.

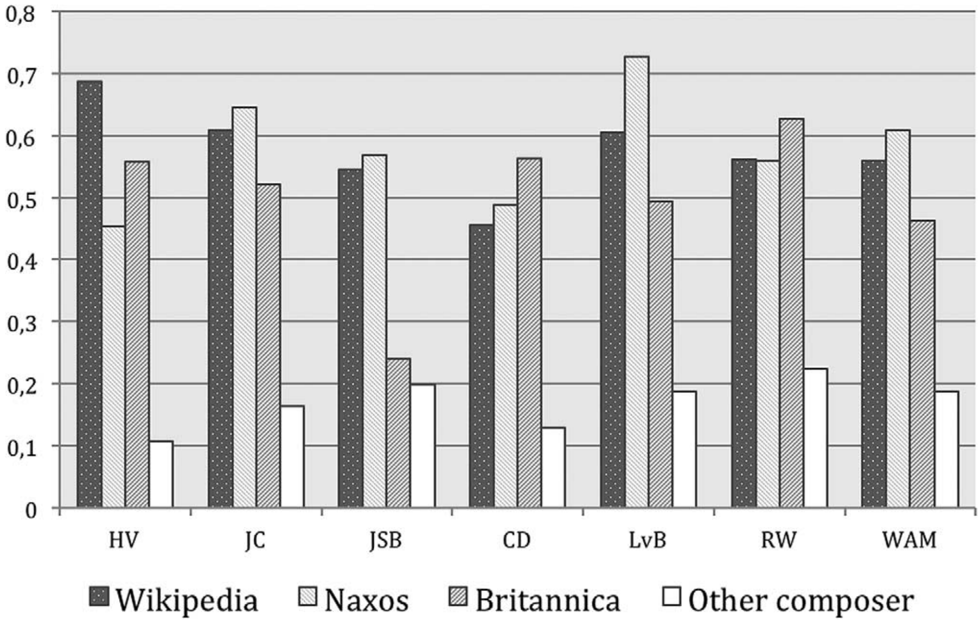


Fig. 2. Similarity between RDF and text files.

similarity. The similarity related to a different composer is in all cases the smallest one. This shows that we are able to detect a text or texts in a collection, which represents the same subject of an RDF file.

Finally, it is worth pointing out that the Wikipedia would be naturally considered as the highest value for this comparison with DBpedia. However, it turns out that only a small fraction from Wikipedia is used in DBpedia. This part is the Infobox that quickly summarizes important points in a structured format. That explains why Wikipedia did not always present the higher similarity between RDF.

6. Related work

Many similarity measurements take into account the context of words to calculate the proximity in textual data. [10] present ways to deal with text similarity in the field of Information Retrieval (IR). By using the IR techniques, [2] apply them to the plagiarism detection. Short segments of text, e.g., Twitter, either do not provide the context or use the non-standard language. Hence, traditional measurements failed in this specific case. The works [1,8] describe alternatives to handle this issue. The studies conducted by [6] evaluate a wide range of the metrics applied to string similarity, and present a guideline on which string metric is more suitable when dealing with ontology alignment systems. [13] survey the state of the art of ontology matching techniques, and present a variety of approaches to explore the topic. [12] study the interaction between the terminological, structural and semantic matching components inside an ontology matching system. Despite the huge volume of writing on similarity measurement, this is restricted to either calculating a similarity between RDF files or textual documents. As far as we know, no researcher has dealt with different source formats, like our approach on calculating the similarity between structured and unstructured data.

7. Conclusion and future works

We have shown an approach to detect a text correspondent to an RDF data calculating the similarity between RDF and text files. The approach is based on a strategy of extracting the literals from an RDF and comparing them to a collection of texts by using the cosine similarity measurement. Our results seem to be a promising avenue to find out how well one format is represented in the other. As future research, we also intend to extract properties from the RDF, which can provide extra information like "birthPlace", "hasAge" etc. and transform them to natural language format to aggregate more contents to the corpus to be compared. Besides, we will deal with negation in natural language, and semantic relatedness to improve our results.

Finally, as our work looks encouraging, in the next steps, it is essential to submit it to a more extensive proof of concept to demonstrate that our work is consistent with different domains and larger datasets.

Acknowledgement

This work was supported by CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) Brazil.

References

[1] F. Abel, Q. Gao, G.-J. Houben and K. Tao, Semantic enrichment of twitter posts for user profile construction on the social web, in: *The Semantic Web: Research and Applications*, Springer, 2011, pp. 375–389.

[2] S.M. Alzahrani, N. Salim and A. Abraham, Understanding plagiarism linguistic patterns, textual features, and detection methods, *Syst. Man, Cybern. Part C Appl. Rev. IEEE Trans.* **42**(2) (2012), 133–149.

[3] T. Berners-Lee, J. Hendler and O. Lassila, The Semantic Web, *Sci. Am.* **284**(5) (2001), 34–43.

[4] C. Bizer, T. Heath, D. Ayers and Y. Raimond, Interlinking open data on the web, in: *Demonstrations Track, 4th European Semantic Web Conference*, Innsbruck, Austria, 2007.

[5] S. Brin, Extracting patterns and relations from the world wide web, in: *The World Wide Web and Databases*, Springer, 1999, pp. 172–183.

[6] M. Cheatham and P. Hitzler, *The Role of String Similarity Metrics for Ontology Alignment*, Dayton, OH, USA, 2013.

[7] D. Duma and E. Klein, Generating natural language from linked data, in: *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, 2013, pp. 83–94.

[8] B. Han and T. Baldwin, Lexical normalisation of short text messages: Makn sens a# twitter, in: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1, 2011, pp. 368–378.

[9] M.A. Hearst, Automatic acquisition of hyponyms from large text corpora, in: *Proceedings of the 14th Conference on Computational Linguistics COLING '92*, Vol. 2, 1992, pp. 23–28.

[10] C.D. Manning, P. Raghavan and H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press, Cambridge, England, 2009, p. 544.

[11] C. Mellish and J.Z. Pan, Natural language directed inference from ontologies, *Artif. Intell.* **172**(10) (2008), 1285–1315.

[12] D. Ngo, Z. Bellahsene and K. Todorov, Opening the black box of ontology matching, in: *The Semantic Web: Semantics and Big Data*, Springer, 2013, pp. 16–30.

[13] P. Shvaiko and J. Euzenat, *Ontology Matching: State of the Art and Future Challenges*, 2012.