Linguistic Summaries of Graph Datasets Using Ontologies: An Application to Semantic Web

Lukasz Strobin(✉) and Adam Niewiadomski

Insitute of Information Technology, Lodz University of Technology,
ul. Wólczańska 215, 90-924 Lodz, Poland
800337@edu.p.lodz.pl, adam.niewiadomski@p.lodz.pl

Abstract. This paper presents a new approach to performing linguistic summaries of graph datasets with the use of ontologies. Linguistic summarization is a well known data mining technique, aimed to discover patterns in data and present them in natural language. So far, this method has been applied only to relational databases. However amount of available graph datasets with associated ontologies is growing fast, hence we have investigated the problem of applying linguistic summaries in this scenario. As our first contribution, we propose to use an ontological class as subject of a summary, showing that its class taxonomy has to be used to properly select objects for summarization. Our second contribution is an extension to a summarizer, by analysis of set of ontological superclasses. We then propose extensions to quality measures T_1 and T_2, measuring informativeness of a summary in the context of ontological class taxonomy. We also show that our approach can create more general summarizations (higher in class taxonomy). We verify our proposals by performing linguistic summarization on Semantic Web, which is a vast distributed graph dataset with several associated ontologies. We conclude the paper with showing the possibilities of future work.

1 Introduction

This paper focuses on performing linguistic summaries on graph datasets with associated ontologies, which has not been attempted before. We propose to rebuild and extend the notion of a subject and a summarizer, by including class taxonomies into problem analysis. For summary subjects we show that all subclasses of a given class have to be analyzed, while for the summarizer - all super classes, which leads to obtaining new information (more general summaries). Secondly, also based on class taxonomies, we propose extensions to quality measures T_1 and T_2.

In this paper we firstly introduce main concepts of linguistic summaries, formerly defined by Yager [1],[2], [3],[4], which are intended for relational databases only. These algorithms provide means of discovering general knowledge and complex patterns in data and presenting it in human-readable quasi-natural sentences. This form of data mining is especially suitable for very large datasets, like Semantic Web. The central notion of our approach is the usage of ontologies, with particular emphasis on class taxonomy. We show how analysis of sub- and superclasses

M. Núñez et al. (Eds.): ICCCI 2015, Part I, LNAI 9329, pp. 380–389, 2015.
DOI: 10.1007/978-3-319-24069-5_36

of a given class can lead to creation of new lingusitic summaries. First of all, in our approach we use an ontological class as a subject of a summary, e.g. 'artist'. Note however that 'is-a' relationship that indicates class membership is a transitive predicate, hence, in a general case, proper selection of summary subjects require analysis of all subclasses of a given subject class (since a 'writer' is also an 'artist'). On the other hand, for creation of summarizers, taking all superclasses of summarizer class (given that this class is a member of an ontology) can lead to creating new (more general) summaries.

This paper is organized as follows. In Section 2 we only remind the reader the main concepts of linguistic summaries. In section 3 we discuss how algorithms for linguistic summaries may be adopted for Semantic Web with the use of ontologies. The exact algorithm of generating summaries for Semantic Web is presented in section 3.5. Proposed algorithm adapts to the dataset, in which the set of summarizers is created dynamically. Section 3.4 shows how T_1 and T_2 quality measures may be extended with class taxonomy. We introduce the notion of summary on different level of generality (Degree of Summarizer Imprecision), depending on the class used in summary. Afterwards, in Section 4 we show the results of an experiment. In the end, in Section 5 we draw the conclusions and show the possibilities of future work.

2 Linguistic Summaries of Relational Databases

This chapter is only meant to remind the reader the main concepts of linguistic summaries, and for in-depth understanding the reader is asked to refer to [1], [2], [3], [4].

Consider the database D. The first form of a linguistic summary is presented by (1):

$$Q \quad P \quad are/have \quad S_j[T] \tag{1}$$

where Q is the *linguistic quantifier*; P is the *subject of the summary* (set of objects represented by the database tuples d_i); S_j is a *property of interest*, the so-called *summarizer* represented by a fuzzy or a crisp set (discrete set in particular).

The crucial part of the algorithm in the sense of Yager is the computation of the degree of truth T. The algorithm is strictly based on Zadeh calculus of linguistically quantified statements, and is computed as:

$$T_1 \ (Q \ P \ are/have \ S_j) \ = \mu_Q(\frac{r}{m}) \tag{2}$$

where

$$r = \sum_{i=1}^{m} \mu_{S_j}(d_i) \tag{3}$$

In typical applications the symbol μ_{S_j} is a membership function of d_i to fuzzy set S_j. However, S_j may also be a discrete set, e.g. $BORN \ IN \ Poland$,

hence $S_j = \{Poland\}$. In this case membership value is given by (4) (this trivial formula is quoted in this paper, because it is a starting point to author's original contribution, see Def. 5).

$$\mu_{S_j}(d_i) = \begin{cases} 1 & \text{if } d_i \in S_j \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

The *degree of truth* as calculated by (2) describes only one of the many aspects of a summary. Quality measure T_2 called *degree of imprecision* describes how imprecise the summarizer is, see (4). The meaning of this quality measure is as follows: the more general the statement is, the higher the value of the imprecision.

$$T2 = 1 - (\prod_{j=1}^{n} in(S_j))^{1/n} \tag{5}$$

where in is the inprecision of a fuzzy set [1].

In this paper we propose an interpretation of the notion of degree of imprecision for ontological classes, see (7) on page 386.

3 Linguistic Summaries of Graph Databases Using Ontologies

The first part of this section introduces a set of concepts and definitions from the field of ontologies. In the remainder of this section author's original contributions are presented - subject selection (using ontological subclasses), extensions of summarizer (using ontological superclasses) and extensions of quality measures T_1 and T_2 (using a complete class taxonomy).

3.1 Definitions Related to Ontological Classes Taxonomy

An ontology is an explicit specification of a conceptualization [7], which defines classes (types), properties of these classes, and taxonomies. In this paper we will denote an ontological class as c.

Consider the fragment of DBPedia ontology shown in figure 1. Say we want to summarize class 'writer'. Belonging to a class is defined by predicate 'rdf:type', and classes in an ontology are linked using predicate 'rdfs:subClassOf' which is transitive (see www.w3.org/TR/rdf-schema for rdf and rdfs reference). As a result, 'rdf:type' is a transitive property with respect to ontological class taxonomy.

Definition 1. *Subclasses of class c are classes, which are directly below class c in a given taxonomy. We denote this set of subclasses by* Sub_c.

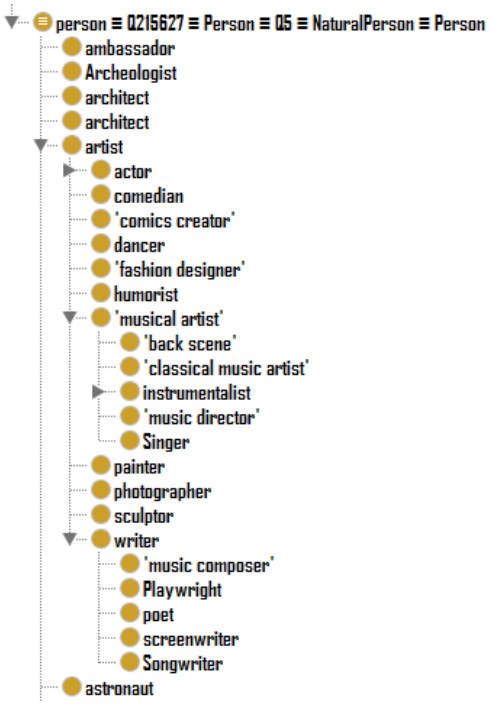

Fig. 1. Fragment of DBPedia ontology, class 'Person'

Example 1. For fig. 1, class *writer* has the following subclasses:
$Sub_{artist} = \{actor, comedian, comics\ \ creator, dancer, fashion\ \ designer, humorist, musical\ \ artist, painter, photographer, sculptor, writer\}$.
Note that $Sub_{writer} \not\subseteq Sub_{artist}$, because we consider only direct subclasses.

Definition 2. *Subclasses of of n-th level of class c are classes, which are separated from class c by not more then n specialization relations. We denote this set by* Sub_c^n.
Note that $Sub_c = Sub_c^1$, $Sub_c^{n-1} \subseteq Sub_c^n$

Definition 3. *The complete set of subclasses of class c (all levels below class c) is denoted by* Sub_c^∞.

Example 2. Let's consider a class *artist* shown in fig. 1. For this class the following sets may be defined:
$Sub_{artist} = Sub_{artist}^1 = \{actor, comedian, 'comicscreator', dancer, 'fashion\ designer', humorist, 'musicalartist', painter, photographer, sculptor, writer\}$
$Sub_{artist}^2 = Sub_{artist}^1 \cup Sub_{actor}^1 \cup Sub_{musical\ \ artist}^1 \cup Sub_{writer}^1$

Analogously to the notion of subclass we define a superclass, and set of superclasses of n-th level - see definitions 1, 2, 3.

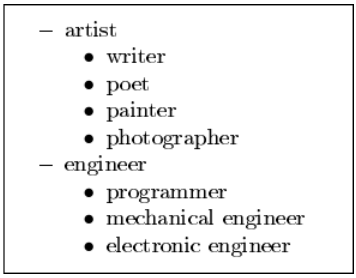

Fig. 2. 'Occupation' value class taxonomy for example

Definition 4. *Superclasses of of n-th level of class c are classes, which are separated from class c by not more then n generalization relation from class c. We denote this set of classes by* Sup_c^n.

3.2 Building Summary Subject Using Class Taxonomy - Including Subclasses

As a subject of a summary, we propose to use an ontological class, with additional consideration of the hierarchy of classes. In a general case, a graph vertex does not specify all classes that is belongs to, but only the most specific one, that is - lowest in class taxonomy. Hence, in order to properly select objects for summarization, also all subclasses of given class have to be selected (see def. 1). Hence, when a linguistic summary of class $'c'$ is created, vertices not only of class $'c'$, but also all vertices of classes Sub_c^∞ have to be selected (see Def. 3), because each member of any of the classes Sub_c^∞ also belongs to class $'c'$.

Example 3. Say we want to create a summary of class 'writer'. We select all vertices of class *writer* but also - classes 'music composer', 'Playwrit', 'poet', 'screenwriter' and 'Songwriter'.

3.3 Building Summarizer Using Class Taxonomy - Including Superclasses

In the proposed method, set of summarizers S is created during selecting objects for summarization, so each triple that has predicate $rdf : type \in Sub_c^\infty$ (see Def. 2). Now for each attribute A_i (that is predicate label) its discrete value set is created based on the values of all retrieved triples. We denote this set of values as X_{A_i}. The attribute A_i may be a graph vertex that has its class $('a_j rdf : type' = c_{A_i})$ that belongs to an ontology - may be the same or different then the subject ontology. In this case, due to transitivity of 'rdf:type' predicate, attribute value a_j also belongs to super classes of class c_{A_i}, hence a set of classes $Sup_{c_{A_i}}^n$ (see def. 4). In this case the set of summarizers is augmented by this set of super classes.

Example 4. Consider a simple ontology 'Occupations' that has the taxonomy shown in figure 2. Say we are creating linguistic summaries of class 'Person', and one of the atributes/predicates is $A_1 =' occupation'$ (A belongs to 'Occupations' ontology). Assume that in the considered dataset the set of values is $X_{Occupation} = \{writer, poet, painter\}$. In this case, in a regular approach, the summarizer $S_{Occupation}$ based on this attribute would have only three possible values $S_{Occupation} = \{writer, poet, painter\}$. However, by taking the set of superclasses of each class in $X_{Occupation}$ and adding to the summarizer set we obtain the following summarizer set: $S_{Occupation} = \{writer, poet, painter, artist\}$. Summarizing on more general attribute value *artist* may lead to extracting new knowledge from the data.

3.4 Ontological Extensions to Quality Measures

T_1 Extended by Class Taxonomy
Recall the summary truth value T_1, given in (1), and the notion of membership function for a discrete summarizer (4). Now consider an ontological class as a summarizer or a qualifier. In this case, the notion of 'being member of a class' may be extended. Since a class $poet \in Sub^{\infty}_{writer}$ (see fig. 1), hence (4) may be extended to have the form as in definition 5.

Definition 5.

$$\mu_c^{ont.}(d_i) = \begin{cases} 1 & if\ d_i \in \{c, S_c^{\infty}\} \\ 0 & otherwise \end{cases} \tag{6}$$

Example 5. Let's consider a summarizer (or a qualifier) 'is writer' and an object, which belongs to a class *poet*. Using a regular approach, so using equation 4 the obtained membership value is 0, while by using an extended approach, equation 6 evaluates to 1.

T_2 Extended by Class Taxonomy
Using (6) for evaluating the membership value of an attribute to a summarizer leads to generating more general summaries, when using classes higher in an ontology. For example, let's imagine a summarizer related to geographical locations, like cities, countries and continents. Assume also that each instance of a class (e.g. $Person$) has a property 'city of birth'. When formula 6 is used, we may create new summaries, extending by the notion of a city to a broader term, like a country. Then we may form a new summary, otherwise not possible (since each attribute specifies only a city), like 'average number of people that are tall are born in Europe'. However, such summaries are less precise - extreme case would be 'most people are born on Earth'. This summary is definitely true, however it is very imprecise.

Hence, we propose an analogous measure to a degree of imprecision for fuzzy sets - we call this notion degree of ontological class imprecision, see definition 6. Proposed formula describes the intuition that the imprecision of the class depends linearly on the number of classes that are below a given class in a taxonomy.

Definition 6. *By degree of ontological class imprecision we call the level of generality of a given class, and evaluated by using (7) in (5).*

$$in^{ont.}(S_j) = \frac{|Sub^{\infty}_{S_j}|}{|Sub^{\infty}_{S_j}| + |Sup^{\infty}_{S_j}|} \tag{7}$$

Example 6. Evaluated degrees of selected ontological classes imprecisions for ontology presented in figure 1 are shown below:

1. $in^{ont.}(Writer) = \frac{5}{5+5} = 0.5$
2. $in^{ont.}(Artist) = \frac{24}{24+4} = 0.85$
3. $in^{ont.}(Person) = \frac{173}{173+3} = 0.98$
4. In the top of the ontology there is a class Thing. For this class the degree of imprecision is equal to 1.

3.5 Generating Linguistic Summaries for Graph Databases - Complete Process with Example

The set of quantifiers Q is known beforehand, as well as the summary subject - an ontological class. We also know the set of ontologies that will be used for summaries - we denote this set of ontologies by T.
For universality of the method, we do not know the attributes, denoted by A, nor their set of values, denoted by X_A. Attributes and their values will be used as summarizers. Exact steps to be followed are listed below.

1. define the ontological class c that will be the subject of the summary
2. generate full set of subclasses for class c Sub^{∞}_{c} (see Def. 2)
3. query the database for objects of classes Sub^{∞}_{c} and their attributes (so vertices that are directly connected to them).
4. based on the queried data we create a set of attributes and their value sets: $A = \langle A_1, X_{A1}\rangle$, $\langle A_2, X_{A2}\rangle$, ... , $\langle A_i, X_{A_i}\rangle$
5. for each attribute A_i we check if it belongs to any considered ontology T, which means to check if it is an ontological class. If so, we take the full set of superclasses of this attribute value $Sup^{\infty}_{A_i}$ (see definition 4). Each of the superclasses may be used to form a more general summary.
6. we create a set of linguistic summaries using found attributes as summarizers - $X_{A_i} \cup Sup^{\infty}_{A_i}$
7. for each qualifier we calculate truth values $T_1 - T_2$

Example 7. Assume that the subject of a linguistic summary is on ontological class writer, and for the summary we will use an ontology GeoNames, which contains information about administrative classification of the world. Hence the set $T = \{GeoNames\}$.

1. $c = writer$
2. $Sub^{\infty}_{writer} = \{'musiccomposer', Playwright, poet, screenwriter, Songwriter\}$ (see fig. 1)

3. query the database for all objects of class *writer* and all subclasses - Sub^{∞}_{writer}
4. $A = \langle 'born', \{Paris, NewYork, Katowice, Amsterdam\}\rangle$, $\langle 'height', \{176cmv, 186cm, 190cm, 166cm\}\rangle$.
5. $'born' \in T$. $Sup^{\infty}_{Paris} \cup Sup^{\infty}_{NewYork} \cup Sup^{\infty}_{Katowice}$ $= \{France, USA, Poland, Europe\}$
6. set of summarizers $S = \{Paris, NewYork, Katowice, Amsterdam, France, USA, Netherlands, Poland, Europe\}$
7. calculation T_1, T_2 and T_{final} - an average of T_1, T_2

4 Application Example - Generating Linguistic Summaries of DBPedia as Part of Semantic Web

DBPedia (see [8], [9]) is an extraction of info boxes from Wikipedia articles into Semantic Web format - Resource Description Framework, RDF (see [10]). In short, RDF is a data format/model composed of triples (subject, predicate and object) that allow easy data integrations. Currently DBPedia contains over 4 milion objects in the main dataset, while it can be easily connected to other data sources using $owl:sameAs$ links available. DBPedia created its own multi-domain ontology which will be used for this experiment, but is also using several others - like subject categories (dcterms), Open Cyc, Wordnet, Freebase, UMBEL and YAGO2. We have implemented our system in Java using jFuzzyLogic [11] and Apache Jena [12] for querying and processing DBPedia (using its SPARQL endpoint). We have used typical triangular definitions of qualifiers.

We have created summaries for a subclass of class Person - class Artists (96300 instances) - see table 1. Due to the nature of the data, there is an unusually large number of summarizers (in comparison to a typical relational database case) - 56. Due to that complexity, including compound summarizers is not directly feasible and we have not included them in the experiment. As can be seen from the table, some interesting patters may be found - for example that about half artists are musicians. A especially interesting summary is summary

Table 1. A small subset of obtained linguistic summaries of DBPedia for class Artists

No.	Summary	T_1	T_2	T_{final}
1	almost none artists have genre jazz[1]	0.86	0.63	0.75
2	almost none artists are born in France[1]	0.92	1.0	0.93
3	small number of artists are born in Europe	0.83	0.11	0.47
4	about half artists are musicians	0.53	0.55	0.54
5	almost none artists are comics creators	0.78	1.0	0.89
6	almost all artists born in Russia are actors	1.0	0.72	0.96
7	small number of artists that play piano are singers	1.0	1.0	1.0
8	about half of artists with genre Soul music are singers	0.98	1.0	0.99
9	about half of artists with genre Soul music are guitarists	0.93	1.0	0.97

number 3 from table 1 - summarizer 'born in Europe' has been obtained by using summarizer generalization described in section 3.3 [1]. Note also the much lower value of T_2 for this summarizer, which indicates that this summary is very general.

5 Conclusions and Future Work

In this paper we have presented a novel approach to linguistic summarization of graph datasets with the use of ontologies. We have rebuild the notion of subject summary by including ontological class taxonomy. Also, we have shown that when an ontological class is used as a summarizer, it is possible and useful also to include the class taxonomy into analysis, since we may obtain summaries that cannot be directly computed with typical approaches, that is more general summaries. Creating such summaries (e.g. summarizing on continental level, while only the information about a country in directly available) leads to finding new dependencies in data. We have also extended the T_1 and T_2 quality measures, also by including class taxonomy into analysis. By quality measure T_2 we are able to determine the informativeness of a summary. We have proven our approach by generating linguistic summaries for a small subset of DBPedia.

Further work may be focused on taking attributes of attributes into account (vertices with distance of 2 edges from summary subjects). As an example consider a subject type 'movie'. A common attribute of this type may be 'director', who also has its properties (like country of birth, age). Another direction of research continuation could be leveraging Linked Data nature - incorporating other ontologies and databases. Since Semantic Web is based on Linked Data, we may also use other ontologies and information sources to create new summaries. For instance, DBPedia is interconnected to DBTune (music database), Eurostat (statistical information), LinkedMDB (movies database), LinkedGeoData (geographical database), GeoSpecies (various information about species) and many others.

Authors of this paper have already conducted research from the area of acquiring knowledge from graph databases, so far focused on path analysis using artificial intelligence [5],[6]. A comprehensive work on extracting knowledge from graph databases using artificial intelligence is being prepared.

References

1. Yager, R.R.: A new approach to the summarization of data. Inf. Sci. **28**(1), 69–86 (1982)
2. Yager, R.R.: Linguistic summaries as a tool for database discovery. In: FQAS, pp. 17–22 (1994)

[1] Since France is not an ontological class, we used a political taxonomy of the world to derive imprecison of summarizer 'France'. For musical genres, like jazz, we used the genre taxonomy given in [13].

3. Yager, R.R., Ford, K.M., Cañas, A.J.: An approach to the linguistic summarization of data. In: Bouchon-Meunier, B., Yager, R.R., Zadeh, L.A. (eds.) Bouchon-Meunier, IPMU. LNCS, vol. 521. Springer, Heidelberg (1990)
4. Yager, R.R.: On linguistic summaries of data. In: Knowledge Discovery in Databases, pp. 347–366 (1991)
5. Strobin, U., Niewiadomski, A.: Evaluating semantic similarity with a new method of path analysis in RDF using genetic algorithms. Journal of Applied Computer Science
6. Strobin, L., Niewiadomski, A.: Recommendations and object discovery in graph databases using path semantic analysis. In: Rutkowski, L., Korytkowski, M., Scherer, R., Tadeusiewicz, R., Zadeh, L.A., Zurada, J.M. (eds.) ICAISC 2014, Part I. LNCS, vol. 8467, pp. 793–804. Springer, Heidelberg (2014)
7. Gruber, T.R.: A translation approach to portable ontology specifications. Knowl. Acquis. **5**(2), 199–220 (1993)
8. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer, C.: DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. Semantic Web Journal (2014)
9. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.G.: DBpedia: a nucleus for a web of open data. In: Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L.J.B., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) ASWC 2007 and ISWC 2007. LNCS, vol. 4825, pp. 722–735. Springer, Heidelberg (2007)
10. Candan, K.S., Liu, H., Suvarna, R.: Resource description framework: metadata and its applications. SIGKDD Explor. Newsl. **3**(1), 6–19 (2001)
11. Cingolani, P., Alcalái-Fdez, J.: jfuzzylogic: a robust and flexible fuzzy-logic inference system language implementation. In: FUZZ-IEEE, pp. 1–8. IEEE (2012)
12. Seaborne, A.: Jena, a Semantic Web Framework, November 2010
13. Garcia, J., Barbedo, A., Lopes, A.: doi:10.1155/2007/64960 research article automatic genre classification of musical signals