A Common Property and Special Property Entity Summarization Approach Based on Statistical Distribution

Yang Li and Liang Zhao

Department of Computer Science and Engineering
East China University of Science and Technology, Shanghai, 200237, China
marine1ly@163.com,252007913@qq.com

Abstract. Combined with previous research, the concept of common property and special property are defined based on the statistical distributions of properties in our paper, and we use the two kinds properties to summarize entities. Common property is a property that the entities under the same type all have. It is the basic property of the entities, and it can help recognize a kind of entities. Special property is a property that just a few entities have, and it can help identify an individual entity among a kind of entities. In addition, we also calculated the importance of property values based on the statistical distributions of property values, so that when a entity has more than one property value with the same property, we choose the triple with more important property value to summarize the entity.

1 Introduction

Thalhammer et al. [1] proposed a method that use the k-nearest neighbors and corresponding properties of a entity to summarize the entity, and the k-nearest neighbors are calculated by the similarity between the entities. The method filters out properties that all entities have, which are called common properties in our paper, and it focuses on the characteristics of entities. Our method takes both similarities and characteristics of the properties in entities into account, and we defined common property and special property from a statistical point of view. Common property can be used to recognize a kind of entities and special property can be used to identify an individual entity among a kind of entities. Those two kind of properties are used to summarize entities. However, the facts of entities (triples) consist of properties and property values, and a entity may have one property with multiple property values. We use the statistical distributions of property values to determine which property value is fit for summarize. The occurrence frequency of the property value reflects its importance, and more important property value is chosen to summarize the entity. Furthermore, there are redundant and useless triples in the facts of a entity. We first filter out those triples and then use our method to do entity summarization.

The main ideas of our method:

– There are three kinds of entities in the LinkedMDB-30 Track, namely, Director, Film, and Actor. The distributions of the occurrence number of each

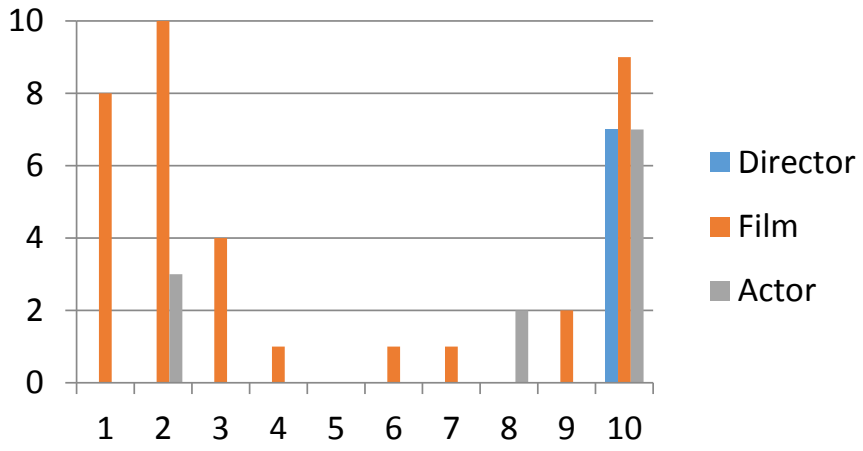

Fig. 1. The distributions of the occurrence number of each property

property under the three classes are shown in Figure 1, and each class contains ten entities. We found that some properties appear in almost all the entities, and we name those properties as common property. Some properties exist only in a handful of entities, and we name those properties as special property. In Figure 1, all properties of the class Director appear in all the ten entities, which indicates that entities of the class Director have only common property.

– Since a entity may have one property with multiple property values, for example, a director may have made more than one film (namely, triples with the same subject and predicate, but different objects), the property value should be considered too. Which property value should be taken to summarize the entity is depend on its importance, and the importance is judged through the dump file (the full data set).

2 Common Property and Special Property Entity Summarization Approach

The processes of our approach consist of four parts: Preprocessing, Property Statistic and Analysis, Property Value Statistic and Analysis, Select Special Triples and Common Triples.

2.1 Preprocessing

Through the observation and analysis of entity files, we find there are many redundant and useless triples that should be deleted, if not, these triples would be selected and the quality of entity summarization would be very low. Here are some examples of triples that need to be deleted. We also introduce the method of deletion.

1. Drop redundant triples. There are some triples which are different in expressing but have the same meaning. For example, in Director entity files exist triples like <director/A> <made> <film/B> and <film/B> <director> <director/A>. Obviously, these two triples mean the same thing, thus we delete one of them. Our deletion method is: if there exists triples A and B, where the subject of A is the object of B, meanwhile, the object of A is the subject of B, then we delete the triple whose object is the entity.
2. Drop useless triples. a) Drop triples whose object value is null. b) Drop triples that contain information of entity id. c) Some triples whose predicate are "page", "rdf-schema#label", or "owl#sameAs" etc. These triples have nothing to do with entity summarization.

2.2 Property Statistic and Analysis

1. As special property and common property mentioned above are based on statistic, we need count each property exists in how many entity files. Next, we regard the property whose occurrence times is more than x as common property and add it in common property candidate list. Likewise, we regard the property whose occurrence times is less than y as special property and add it in special property candidate list. Here x and y are the threshold parameters available for setting.
2. After generating the two candidate lists, we need to calculate the degree of commonness and the degree of specialness and rank them from high to low separately. We record each propertys occurrence times in each entity file n, occurrence times in the 10 entity files N, and occurrence times in dump file N_{dump}. We use the following equation to calculate the degree:

$$SpecialDegree = \frac{n}{N} + \frac{n}{N_{dump}} * 1000 \tag{1}$$

$$CommonDegree = \frac{N_{dump}}{N} \tag{2}$$

Equation (1) is used to calculate the special degree. $\frac{n}{N}$ is the special degree of a special property in the 10 entity files and $\frac{n}{N_{dump}}$ is the special degree of a special property in the dump file. As the dump file is huge, the value of $\frac{n}{N_{dump}}$ is quite small, after analyzing each special propertys $\frac{n}{N_{dump}}$, we find it varies from 0.0001 to 0.001, so we multiple a balance parameter 1000 to scale its influence. Equation (2) is used to calculate the common degree. We do not take $\frac{N}{n}$ into account where $\frac{N}{n}$ is the common degree of a property in the 10 entity files. Because the range of $\frac{N}{n}$ is about 10 and the value of $\frac{N_{dump}}{n}$ varies from 3000 to 20000. We calculate propertys commonness in dump file only, because the information in the 30 entities is too small, which making their commonness not obvious. According to the degree, we give the two candidate lists a rank from high to low separately.

2.3 Property Value Statistic and Analysis

After obtaining special properties and common properties, the analysis and statistics of property values are also needed. There exists such a situation that one property has multiple property values in one entity. For example, a director has a property "made", which is used to describe that the director made a film. Obviously, one director could make many films. When we use property "made" to summarize director entity, a strategy to determine which film to choose is required. We propose a method as follows:

Extract each films information from dump file. In dump file, most films have a property "performance" but the values are different. We hypothesis the values of property "performance" is the film entitys score, and high score implies high importance. We choose the most important film to summarize director entity. Similarly, when summarizing film entities, we may use "actor" property. We also extract actor information from dump file. We hypothesis that the more films the actor participates, the more important the actor will be. The actors score is the number of films he or she participates.

For actor entities and director entities, they both have two kinds of type. Take actor entity as example, one of its type is "person" and the other is "actor", obviously, "actor" is a subclass of "person" and the sub class is more accurate when summarizing entity. There is a truth that in dump file, the amount of triples which describe superclass is bigger than that which describe subclass. Thus when an entity has two types, we select the one which has less appear times in dump file.

2.4 Select Special Triples and Common Triples

Select z triples which contain special property and 5-z triples which contain common property to summarize entity, where z is a parameter available for setting. As common property is much more than special property and not all of entity files meet the requirement of z special properties (e.g. director entity), z is just a maximize value. If there are only t (t<z) special triples, then 5-t common triples would be selected. The special properties and common properties are selected from the two ranked lists generated in 2.2 respectively. When processing the property with multiple values (i.e. actors or films), we select candidates from high to low score mentioned in 2.3.

References

1. Thalhammer, A., Toma, I., Roavalverde, A., Fensel, D.: Leveraging usage data for linked data movie entity summarization. Computer Science - Artificial Intelligence (2012)