Coarse to Fine: Diffusing Categories in Wikipedia

Pengshan Cai[1], Yansong Feng[2], Yantao Jia[1], Yuanzhuo Wang[1], Xiaolong Jin[1], Xueqi Cheng[1]
[1]CAS Key Laboratory of Network Data Science and Technology,
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[2]Institute of Computer Science & Technology, Peking University, Beijing, China
caipengshan14@mails.ucas.ac.cn; fengyansong@pku.edu.cn;
{jiayantao, wangyuanzhuo, jinxiaolong, cxq}@ict.ac.cn

ABSTRACT

Automatic taxonomy construction aims to build a categorization system without human efforts. Traditional textual pattern based methods extract hyponymy relation in raw texts. However, these methods usually yield low precision and recall. In this paper, we propose a method to automatically find diffusing attributes to a category from Wikipedia infoboxes. We use the diffusing attribute to diffuse a coarse-grained category into several fine-grained subcategories and generate a finer-grained taxonomy. Experiments show our method can find proper diffusing attributes to categories across various domains.

Keywords

Wikipedia; Category; Diffusing Attribute; Taxonomy

1. INTRODUCTION

Wikipedia has a tree-structured taxonomy[1], in which both subcategories and entities are linked below their parent categories. As new concepts continuously emerge, Wikipedia encourages its maintainers to update its taxonomy frequently. However, this work takes enormous human efforts, from 2012 to 2014, 258,483 new categories were manually added to Wikipedia. Even so, some categories (especially those maintained by few maintainers) are still coarse-grained and should be further diffused into several finer-grained subcategories. E.g., the category *Chinese Opera* could be further diffused into *Beijing Opera*, *Kunqu Opera*, etc.

Previous works such as [**?**] attempt to automate taxonomy construction using textual patterns . However, many hyponymy relations were not explicitly expressed in the text and can not be extracted by textual patterns. As a result, these methods usually achieve low precision and recall. In fact, Wikipedia suggests another way of generating new subcategories from an existing category, i.e. to diffuse a category into subcategories by a diffusing attribute. For example, by attribute *OS*, we diffuse the category *Mobile Phones* into *Mobile Phones of IOS, Mobile Phones of Android*, etc.

[1]https://en.wikipedia.org/wiki/Wikipedia:Categorization

WWW 2017 Companion, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4913-0/17/04.
http://dx.doi.org/10.1145/3038912.3038914

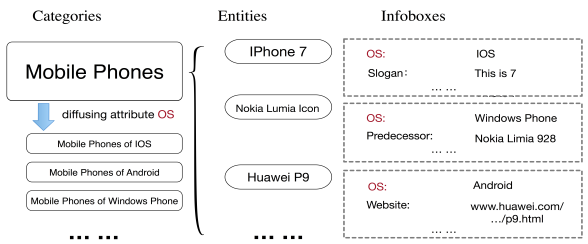

Figure 1: An illustration of using attribute *OS* to diffuse the category *Mobile Phones*

In this paper, we propose an unsupervised method to automatically discover diffusing attributes to a Wikipedia category. Take *Mobile Phones* in Figure 1 as an example, we collect all entities under this category, including *iPhone 7, Huawei P9*, etc. For each entity, we collect all its attributes and corresponding values in its infobox. Obviously, some attributes could be used as diffusing attributes to the category, e.g. *OS*, etc. While others can not, e.g., *Slogan, Predecessor*, etc. We score each attribute using a scoring function and rank them in descending order, the higher an attribute ranks, the more likely it could be used as a diffusing attribute to the category. Finally, we diffuse the category into subcategories. Experiments on Wikipedia categories of various domains proved our method effective.

2. FINDING DIFFUSING ATTRIBUTES

We assume that a diffusing attribute should meet the following requirements. **Relatedness**: the diffusing attribute should be closely related to the category to be diffused; **Effectiveness**: The diffusion should enhance users' navigation instead of offering them redundant information.

1)Relatedness. We note that, the more entities under a category C are related to an attribute a (i.e. More entities under category C hold values on attribute a), the more proper a is a to diffuse C. For instance, almost all entities under the category *NBA Basketball Player* hold values (e.g. *Forward, Center, Guard* etc.) on the attribute *Position*, which is a proper diffusing attribute to the category *NBA Basketball Player*. On the other hand, the less entities under C are related to a, the less proper is a to diffuse C. For example, Yao Ming, a famous NBA basketball player, is also a voice actor. However, most other entities under the cate-

gory *Voice Actor* do not hold value on Yao Ming's attribute *Position* in their infoboxes, and the attribute *Position* is not proper to diffuse the category *Voice Actor*.

2)Effectiveness. The purpose of the categorization system is to enhance navigation, thus, two reasons may lead to an improper category diffusion: 1) Each subcategory contains very few entities. For example, the attribute *Logo* is not suitable to diffuse the category *Software Companies*, as each company has a unique logo, diffusing the category *Software Companies* by the attribute *Logo* would only offer redundant information. 2) The diffusion cause most entities under the parent category fall into just one subcategory. This could lead to an unbalanced and deeper taxonomy which prolongs users' navigation. For example, in the category *LG Electronics Mobile Phones*, most entities hold value *LG Electronics* on attribute *Developer*, while *Nexus 5* is the only entity that holds *Google and LG Electronics* on this attribute. Using *Developer* as a diffusing attribute would generate two subcategories, one contains only *Nexus 5*, the other contains all other entities under the parent category.

Inspired by the C4.5 algorithm [**?**] which uses information gain ratio to measure how proper a feature can be used for classification, we define diffusing ratio $DR(a, C)$ to to measure how proper is an attribute a to diffuse a category C. Formally,

$$DR(a, C) = \frac{IV(a, C)}{|V|} = \frac{\sum_{v=1}^{|V|} \frac{|E^v|}{|E|} log_2 \frac{|E^v|}{|E|}}{|V|} \quad (1)$$

where $|V|$ is the number of possible values a can hold, $|E|$ is the number of entities under category C and $|E^v|$ is the number of entities under C that hold value v on attribute a. Here, $IV(a, C)$ is called the intrinsic value of a to C.

Let us check the validation of formula (1) from aspects of relatedness and effectiveness: (i) If more entities under category C hold values on attribute a, DR would be greater. This shows diffusing ratio favors attributes more closely related to the category to be diffused. (ii) When IV is fixed and $|V|$ increases, DR would decrease. This indicates diffusing ratio favors attributes that hold fewer values. (iii). When $|V|$ is fixed, if entities in the coarse category are more evenly diffused into subcategories, IV would be greater. This indicates DR favors attributes that would generate a shallow and balanced taxonomy. II and III shows diffusing ratio is in accordance with the requirement of Effectiveness.

3. EXPERIMENTS

We run experiments on 24 random Wikipedia categories[2]. We compare our method with four baselines: 1) AF ranks attributes by its frequency in Wikipedia. 2) REL ranks attributes by its relatedness to the parent category, i.e. $REL(a, C) = (\sum_{v=1}^{|V|} |E^v|)/|E|$. 3) $IV(a, C)$ in formula (1), it favors attributes related to the category and diffuse the category into similarly sized subcategories. 4) $\text{REL}_{\text{norm}} = REL(a, C)/|\text{V}|$, it favors attributes both closely related to the parent category and have fewer possible values.

We run each method on the 24 categories. Attributes ranked in top 5 are selected as diffusing attributes to each category. As no ground truth is available, we ask three volunteers to evaluate the quality of diffusing attributes each method selects. (Evaluation criteria: 3 points - *a correct diffusing attribute*, 1 point - *hard to determine*, 0 point - *not a diffusing attribute.*) We report the average score, NDCG@3 and NDCG@5 of each method in Table 1. In Table 2, we list a few example categories and their diffusing attributes found by our method.

[2]Please refer to https://github.com/OpenKN-ICT/coarse2fine for source code and experiment details

Method	AvgScore	NDCG@3	NDCG@5
AF	1.20	0.43	0.39
REL	2.42	0.81	0.79
IV	2.20	0.73	0.71
REL_{norm}	2.61	0.93	0.88
DR	**2.83**	**0.95**	**0.94**

Table 1: Human evaluation results

Parent Category	Diffusing Attributes
Propeller_Aircrafts	propeller, nationality, manufacturer
Mobile_Phones	input, os, developer
Sculptures	artist, material, city
Snack_Foods	country, main_ingredient, type

Table 2: Categories and their diffusing attributes

We observe that: 1) Our DR outperforms all baselines in all metrics. Specially, it achieves 2.83/3.00 in AvgScore. This indicates our DR can effectively find diffusing attributes across various domains. On the contrary, AF, which does not consider relatedness and effectiveness when selecting diffusing attribute, performs the worst. This implies relatedness and effectiveness are viable metrics for selecting diffusing attributes to categories. 2)Compared to IV, DR shows much improvement. Because by dividing $|V|$, DR avoids diffusion that generates many small subcategories (e.g. diffusing *Software Companies* by *Logo*). 3) REL_{norm} performs close to our DR. We note that the measurement of REL_{norm} is close to the requirements of relatedness and effectiveness. However, DR has the advantage of favoring attributes which help generate a shallow taxonomy. Although this characteristic may not be appreciated by human evaluators, it is important because a shallow and balanced taxonomy can help users find their information more efficiently.

4. CONCLUSION

In this paper, we propose a method to automatically select diffusing attributes to a Wikipedia category, the selected diffusing attribute can diffuse a coarse-grained Wikipedia category into several fine-grained subcategories. Experiments demonstrate the effectiveness of our method.

5. ACKNOWLEDGMENTS

This work is supported by National Key Research and Development Program of China (No. 2016YFB1000902) and National Grand Fundamental Research 973 Program of China (No. 2014CB340401),National Natural Science Foundation of China (No. 61572469, 61402442, 91646120, 61572473, 61402022).

6. REFERENCES

[1] M. A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *COLING 1992*.

[2] J. R. Quinlan. *C4. 5: programs for machine learning*. Elsevier, 2014.