Evaluation of an Ontology Summarization Approach

Ning Li
Knowledge Media Institute
The Open University
Milton Keynes
United Kingdom, MK7 6AA
n.li@open.ac.uk

Enrico Motta
Knowledge Media Institute
The Open University
Milton Keynes
United Kingdom, MK7 6AA
e.motta@open.ac.uk

Mathieu d'Aquin
Knowledge Media Institute
The Open University
Milton Keynes
United Kingdom, MK7 6AA
m.daquin@open.ac.uk

Zdenek Zdrahal
Knowledge Media Institute
The Open University
Milton Keynes
United Kingdom, MK7 6AA
z.zdrahal@open.ac.uk

ABSTRACT

Ontology summarization is a very useful technique to help users making sense of ontologies quickly. We have developed a summarization approach that linearly combines a number of criteria, drawn from cognitive science, network topology, and lexical statistics to produce ontology summaries [1]. Motivated by our later findings that the approach, in its current form, binds the criteria so tightly that hinders its flexible and optimal usage in different scenarios, this work presents an objective evaluation of this approach. This is not just a supplement to the subjective evaluation already done, but with a more important goal to evaluate the impact and find the ranking of importance for each criterion.

1. INTRODUCTION

With the number and size of ontology increasing as well as complexity of ontology taxonomy, Ontology summarization has been recognized as an important tool to facilitate ontology understanding in order to support tasks like ontology reuse. We developed such an ontology summarization approach [1], called Key Concept Extraction (KCE). It uses a number of criteria drawn from cognitive science, network topology, and lexical statistics to extract key concepts, which are believed to be most reprehensive of the ontology. Ontology summaries produced in this way have been shown to correlate significantly with the ones generated by human experts, referred to as "ground truth". This approach has been used as the basis for a novel ontology navigation and visualization tool, called KC-Viz[1], and also to provide summary view for online ontology sharing and reusing system Cupboard[2].

Though good results were produced in the approach to Key Concept Extraction, the algorithm, in its current form, have limitations on matters, like time constrains, when used in different scenarios. With only subjective evaluation on the final summarization results that is an accumulated effect of all the criteria used in the algorithm, it is not possible to separate the impact of each criterion on and its contribution to making results as close as possible to experts' opinions. Hence, there is a need to evaluate each criterion separately in a comparative manner. Also, a closer look into how they relate to ontology features would be useful. In addition, it provides indicative view of how to improve the overall performance of KCE, by giving optimal weights to each criterion. These weights were only derived empirically in [1], where a comprehensive analysis of the algorithms and associated performances had not been realized.

We start with a review of the current algorithm for Key Concept Extraction in Section2. We will then focus on the main contributions of this paper, that is to objectively evaluate the criteria comparatively in Section 3. In Section 4, we analyze and discuss the evaluation results, and Section 5 concludes the paper.

2. THE KCE ALGORITHMS

In [1], a number of criteria were considered, and correspondingly a number of algorithms were developed, to identify key concepts of an ontology. In particular, the notion of ***natural category***, drawn from cognitive studies, was used to identify concepts that are information-rich in a psycho-linguistic sense. This notion was realized by two operational measures: *name simplicity* which favors concepts that are labeled with simple names, such as *Vegetation* while penalizing compounds such as *ExoticVegetation*; and *basic level* which measures how "central" a concept is in the taxonomy of the ontology, i.e. how many times it appears in the middle of a path from the root to a leaf of the branch that contains the concept. Two other criteria were drawn from the topology of an ontology: the notion of ***density*** highlights concepts that are richly characterized with properties and taxonomic relationships, such as *isA* or *typeof*; while the notion of ***coverage*** aims to ensure that no important part of the ontology is neglected. Lastly, the notion of ***popularity***, drawn from lexical statistics, was introduced to indentify concepts that are commonly used. The ***density*** and ***popularity*** criteria were both decomposed into two sub-criteria, *global* and *local density*, and *global* and *local popularity* respectively. While the global measures are normalized with respect to all the concepts in the ontology, the local ones consider the relative density or popularity of a concept with respect to its surrounding concepts. The aim is to ensure that "locally significant" concepts get a higher score, even though they may not rank too highly with respect to global measures. Each of these seven criteria produces a score for each concept in the ontology and the final score assigned to a concept is a weighted summation of the scores resulted from individual criterion.

3. EVALUATION OF KCE ALGORITHMS

Kendall's tau Statisitcs [2] (abbreviated as *tau*) is often used to measure the agreements between two measured quantities. In specific, it is a measure of rank correlation, that is, the similarity of the orderings of the data when ranked by each of the quantities. It has been used in the evaluation of text summarization [3] as well as an RDF-sentence-based ontology summarization [4]. Here, we use *tau* to find the correlation between the score vector (one per ontology and the length of vector equals the number of concepts in each ontology), produced by each criterion, with human experts' "ground truth" score vector. Eight people with experiences on ontology engineering were asked to select up to 20 key concepts for each ontology. The score vector for each criterion is obtained by running the corresponding algorithm, and

[1] http://neon-toolkit.org/wiki/KC-Viz

[2] http://kmi-web06.open.ac.uk:8081/cupboard/

that of "ground truth" is obtained by counting the experts' votes on each concept and then normalizing the result with respect to the total number of votes being cast in the whole ontology. We still use the same four ontologies *biosphere, music, financial,* and *aktors portal* (see [1]), to find the *tau* scores and their average. Table 1 shows the results. Each entry in this table is the correlation between the criterion score vector and "ground truth" score vector. An average over all test ontologies is listed in the bottom row.

Table 1. Algorithms and experts agreement measured by *tau*

	Coverage	**Global Density**	**Local Density**	**Basic Level**	**Name Simplicity**	**Global Popularity**	**Local Popularity**
Biosphere	0.140	0.454	0.449	0.388	0.111	0.300	0.091
Financial	0.053	0.539	0.547	0.448	0.464	0.430	0.310
Music	0.272	0.308	0.307	0.367	-0.048	0.085	-0.019
Aktors portal	0.241	0.378	0.355	0.401	0.136	0.114	0.055
Average	0.177	0.420	0.415	0.401	0.166	0.232	0.109

The resulted *tau* score does not reflect the precise contributions of each criterion, rather it is often a relative comparison among the criteria. Increasing values imply increasing agreement between the two sets of rankings. If the rankings are completely independent and uncorrelated, the coefficient then has value zero on average. In our case, the higher the score is, the more correlations between the corresponding criterion's score with experts' score and hence the more agreements between their choices of key concepts. Also, it must be emphasized that the scores are most meaningful when considered per ontology. For example, it is not expected to compare the *global popularity* score of *financial* ontology with *global density* of *music* ontology, nor to compare the *global popularity* score of *financial* ontology with *global popularity* of *music* ontology even because two ontologies may have very different features which, as will be analyzed next, affect the definite values of the *tau* score. Only the comparison among different criteria within one ontology indicates the importance of each criterion. Obviously, if one criterion consistently produces higher scores than the other criteria cross all ontologeis, it is reasonable to believe that it is a more important criterion. The average scores listed in the bottom row provide such an indication.

4. ANALYSISES AND DISCUSSIONS

From the results, we can see that the criteria *global density*, *local density* and *basic level*, show consistent high agreements with "ground truth" across all onotlogies with a similar order of rankings, which indicates that human experts also have their attentions on those corresponding features of ontology. While other criteria *coverage*, *name simplicity*, *global popularity, local popularity* show consistent less importance. But the order of rankings among them varies slightly across four ontologies. Though the average score at the bottom row provides the most comprehensive indication of the importance of each criterion, a closer look into those variations could provide a profound insight into the impact of each criterion on ontologies with distinctive features. For example, the ranking of *name simplicity* is lower than *global popularity* in *biosphere* ontology but higher in *financial* ontology. So, why, in another word, *name simplicity* is less important than *global popularity* in *biosphere* ontology but more important in *financial* ontology. Firstly, by looking at what's typically contained in *biosphere* ontology, we know that a majority of the terms are simple names instead of compounds, and also a high percentage of the terms are not popular words. With "ground truth" containing key concepts like *Animal, Bird, Fungi, Insect, Mammal, MarineAnimal* etc., all with very popular names and only one is compound, it is obvious that the impact of *name simplicity* criterion is less prominent than that of *global popularity* in making the summarization results correlating with "ground truth". While for *financial* ontology, a majority of the terms are labeled with popular words and it is often the case that a simple name is franchised by many compound names, With "ground truth", e.g. *Bank, Bond, Broker, Capital, Contract, Dealer, Financial_Market* etc. containing one compound name only, it is not surprising that *name simplicity* may impose a larger impact than *global popularity* on the summarization results in making them correlate with "ground truth" more.

Though lack of comparison value, the definite values for the scores of different criteria are worth looking into. For example, the *global popularity* scores of both *biosphere* and *financial* are pretty high. This in fact reinforces the subjective evaluation in the original work [1]. The initial design of the algorithm did not have *popularity* criterion and the resulted summaries had very low levels of agreement with the "ground truth". When adding *popularity* as an additional criterion to the existing criteria stack, the resulted summaries were all improved significantly, with ontology *biosphere* and *financial* being improved more by 167% and 100% respectively than ontology *music,* and *aktors portal* which had improvement ratios of 50% and 20% respectively (see [1]). Hence, it is not so surprising to see *popularity* criterion has relatively higher *tau* scores for *biosphere* and *financial* than the other two ontologies.

5. CONCLUSIONS

This paper provides an objective evaluation of the Key Concept Extraction algorithms used in an ontology summarization approach. The evaluation results provide a basis to judge the importance of each individual criterion being used. It helps to decide which criterion is prioritized to use or given more weights when such a decision is required in certain use case scenarios.

6. REFERENCES

[1] Peroni, S., Motta, E., d'Aquin, M. 2008. Identifying Key Concepts in an Ontology Through the Integration of Cognitive Principles with Statistical and Topological Measures. In *3rd Asian Semantic Web Conference*, Bangkok, Thailand.

[2] Sheskin, D.J. 1997. *Handbook of Parametric and Nonparametric Statistical Procedures*. CRC Press.

[3] Donaway, R.L., Drummey, K.W., Mather, L.A. 2000. A Comparison of Rankings Produced by Summarization Evaluation Measures. In *ANLP/NAACL Workshop on Automatic Summarization*, pp 69–78.

[4] Zhang, X., Cheng, G., Qu, Y. 2007. Ontology Summarization Based on RDF Sentence Graph. In *16th International World Wide Web Conference (WWW2007)*, Banff, Alberta, Canada, May 8-12.