On Connectivity of Automatically Extracted Keyphrase Graph of Object Descriptions in "Open Karelia" Tourism Information System

Ksenia Lagutina, Eldar Mamedov, Nadezhda Lagutina, Ilya Paramonov
P.G. Demidov Yaroslavl State University, Yaroslavl, Russia
lagutinakv@mail.ru, eldar.mamedov@e-werest.org, lagutinans@gmail.com, Ilya.Paramonov@fruct.org

***Abstract*—In this paper we started to solve the task of tag extraction from object descriptions of the Open Karelia tourism information system so that the objects linked by tags formed highly connected graph. We experimented with the current Open Karelia algorithm and several well-known methods for keyphrase extraction (namely TextRank, Topical PageRank, Kea, Maui) on the Open Karelia database for compliance with the stated requirements and based on the results outlined some possible improvements for keyphrase extraction procedure.**

Open Karelia is an information system about Russian and Finnish museums of the Karelian region. It allows any user to easily access all museums content like their history, exhibitions, excursions, work schedule, ticket pricing, contacts and so on. One of core features of the system is provision of information about objects stored in the museums. The system contains various information about each object including its description, category, class, type, dating, location on the map and so on [1].

Each Open Karelia object is accompanied with a list of relevant tags where tags are actually keyphrases extracted from object's description. Unfortunately, the current tag extraction algorithm has several weaknesses: sometimes the extracted tags poorly characterize the object and most of tags belong only to a small amount of objects. It leads to situations when Open Karelia users starting the site exploring from one object cannot find any other object using links between their tags.

This work in progress is aimed at improvement of the keyphrase extraction algorithm to automatically elicit tags from object descriptions stored in the Open Karelia database so that they meet the following requirements:

- the extracted keyphrases should characterize the object;
- number of objects that correspond to a concrete keyphrase should be as much as possible;
- all objects should be connected to each other by keyphrases and form a highly connected graph with texts as vertices and common keyphrases between texts as edges.

The Open Karelia database contains 986 texts that describe Karelian cultural objects, museum exhibits and tourist attractions. Each text has information about a concrete tourist object. The specificity of the Open Karelia texts lies in the fact that they contains a lot of dates, geographical, personal and other proper names that can be good candidates of keyphrases.

Automatic keyphrase extraction is a search of words and phrases that describe main topics of the corresponding text. Existing methods for automatic keyphrase extraction can be divided into two categories: supervised and unsupervised approaches [2]. For each category we chose one well-known method and one its extension. So we use four algorithms for experiments: TextRank, Topical PageRank, Kea and Maui.

TextRank [3] is an unsupervised graph-based approach for keyphrase extraction. The basic idea behind this approach is to build a graph from the input document with candidate keyphrases as nodes and rank the nodes according to their importance using a special graph-based ranking method. To connect nodes in the graph it uses the co-occurrence relation between words. The idea of the ranking method is that a node is important if there are other important nodes pointing to it. This can be regarded as voting or recommendation among nodes.

Topical PageRank [4] is another unsupervised graph-based approach that is similar to the TextRank. Just like TextRank it is based on building a graph of candidate keyphrases and rank the nodes according to their importance but with some improvements of ranking method. The main difference of these two approaches is what Topical PageRank considers the topics of words and document in the graph ranking stage that ensures that the extracted keyphrases cover the main topics of the document. It runs TextRank multiple times for a document, once for each of its topics induced by a Latent Dirichlet Allocation model.

Kea [5] is a supervised approach for automatic keyphrase extraction from text documents. At first, it finds candidate keyphrases using lexical methods and calculates several statistical features for each candidate. Then Kea builds a prediction model by training on documents with manually extracted keyphrases and finally applies the Naive Bayes algorithm to determine keyphrases for each document. Kea can also use a thesaurus for keyphrase extraction from a controlled vocabulary.

The Maui approach is based on Kea and uses the same schema for automatic keyphrase extraction. The difference between them consists in the fact that Maui calculates more features for candidate keyphrases and applies bagged decision tree method instead of the Naive Bayes algorithm [6].

For the last two algorithms we used the RuThes thesaurus (http://www.labinform.ru/pub/ruthes/) containing 115 000 Russian phrases and relations between them.

ISSN 2305-7254

TABLE I. CONNECTED COMPONENTS OF TEXTS GRAPHS

Algorithm	Number of connected components
TextRank	41, including 40 isolated vertices
Topical PageRank	40, including 39 isolated vertices
Kea	158, including 157 isolated vertices
Maui	1
Current algorithm	12, including 11 isolated vertices

One of requirement for extracted keyphrases in Open Karelia is what all texts should be connected to each other. We consider what two texts are connected if they have a common keyphrase. To estimate texts connectivity for all algorithms outcomes we did the following. Firstly, we run all algorithms on all 986 texts to extract keyphrases. Then we built a graph where each vertex correspond to one text and connect vertices by an edge if corresponding texts have a common keyphrase. After that we calculated the number of connected components of the graphs. The results are shown in Table I.

The results show that Maui is the best for the texts connectivity requirement as it was capable to connect all the texts in the database. All the remaining algorithms left some of texts isolated from the others.

To estimate the number of texts that match a concrete keyphrase we run all algorithms on 100 texts from the Open Karelia database. Then we counted the number of corresponding texts for each keyphrase and calculated several statistical characteristics for each distribution: the minimum number of corresponding texts, the maxinum number of corresponding texts and the median of a corresponding texts numbers distribution. The results of this experiment showed that the number of texts per keyphrase is very low for all the considered algorithms.

Besides, to evaluate quality of keyphrase extraction by the algorithms we chose 100 texts from Open Karelia database, which have keyphrase sets extracted by an expert. These keyphrases were compared with the algorithms outcomes in order to evaluate the quality of the algorithms.

In summary, the results of our experiments showed that standard algorithms do not provide high connectivity of texts and a sufficiently large number of texts corresponds to a single keyphrase. We suggest the following ways to solve this issue:

- extend existing algorithms by a block that extracts proper names as keyphrases;
- improve existing algorithms so that they could use a thesaurus more efficiently;
- automate the extension of a keyword involving hypernyme/hyponyme relations between the terms.

These considerations are the subject for further research.

The research was supported by the grant of the President of Russian Federation for state support of young Russian scientists (project MK-5456.2016.9).

REFERENCES

[1] I. Paramonov, E. Mamedov, S. Averkiev, I. Shchitov, K. Krinkin, and M. Zaslavskiy, "Open Karelia — an informational portal for museums," in *Proceedings of the 17th Conference of Open Innovations Association FRUCT. Yaroslavl, Russia, 20-24 April 2015*. IEEE, 2015, p. 331.

[2] K. S. Hasan and V. Ng, "Automatic keyphrase extraction: A survey of the state of the art," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2014, pp. 1262–1273.

[3] R. Mihalcea and P. Tarau, "TextRank: Bringing order into texts," Association for Computational Linguistics, 2004.

[4] Z. Liu, W. Huang, Y. Zheng, and M. Sun, "Automatic keyphrase extraction via topic decomposition," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2010, pp. 366–376.

[5] I. H. Witten, G. W. Paynter, E. Frank, C. Gutwin, and C. G. Nevill-Manning, "Kea: Practical automatic keyphrase extraction," in *Proceedings of the Fourth ACM Conference on Digital Libraries*. ACM, 1999, pp. 254–255.

[6] O. Medelyan, E. Frank, and I. H. Witten, "Human-competitive tagging using automatic keyphrase extraction," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, vol. 3. Association for Computational Linguistics, 2009, pp. 1318–1327.