Clustering web search results using Wikipedia resource

C. TRAN, A. AMELJAŃCZYK

chung.tran@wat.edu.pl, andrzej.ameljanczyk@wat.edu.pl

Military University of Technology, Faculty of Cybernetics
Institute of Computer and Information Systems
Kaliskiego Str. 2, 00-908 Warsaw, Poland

The paper presents a proposal of a new method for clustering search results. The method uses an external knowledge resource, which can be, for example, Wikipedia. Wikipedia – the largest encyclopedia, is a free and popular knowledge resource which is used to extract topics from short texts. Similarities between documents are calculated based on the similarities between these topics. After that, affinity propagation clustering algorithm is employed to cluster web search results. Proposed method is tested by AMBIENT dataset and evaluated within the experimental framework provided by a SemEval-2013 task. The paper also suggests new method to compare global performance of algorithms using multi – criteria analysis.

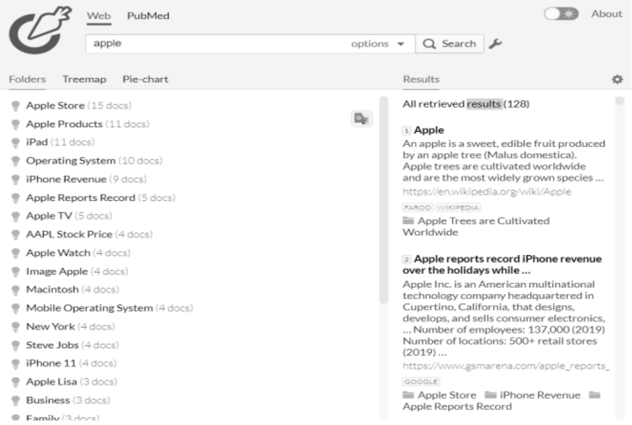
Keywords: affinity propagation, similarity of documents, multi-criteria assessment, Web search results, Wikipedia.

DOI: 10.5604/01.3001.0014.4437

1. Introduction

Every online search engine can give us millions of links to webpages for a single query almost immediately. Similar search results can emerge on the top of a single ordered list sorted by matching scores against issued query. Users must spend a lot of time or even cannot reach desired document if it is far from the beginning of the list. To help user browse a large set of search results with less effort, search results clustering systems reorganize similar search results into smaller groups called clusters before displaying to user. Preparation input, especially calculating similarity of documents for clustering algorithm is a very important of this action. Each search result returned by search engine usually contains title and link to document, a short text fragment (aka snippet) related to search query. Snippet contains sentences which have keywords that appear in the query. The title and snippet are, essentially, representation of search result with user. Usually, these results are ranked according to calculated scores by system and displayed to user about 10 results per page. However, most of the queries are short, ambiguous, polysemy and user checks several first pages of results only [14]. Therefore, there is less chance for webpage at the end of ranked list is visited by user and on the other hand, user may not reach to desired result while top positions in ranked list are populated by similar results. Search Results Clustering (SRC) is one among methods to help users get a complementary view to results list and find relevant document while spending less time. Based on this new view, user can acquire new knowledge about issued query or refine the query.

Given a ranked list returned by search engine, a SRC system groups search results based on their similarity into labeled clusters, also called categories. For example, the term "apple" can refer to a kind of fruit or a technology product or even a company, so a SRC system is expected to group results related to agriculture in one category and results related to technology in another category and so forth. An example is illustrated below.



All retrieved results (128)

1 **Apple**
An apple is a sweet, edible fruit produced by an apple tree (Malus domestica). Apple trees are cultivated worldwide and are the most widely grown species ...
https://en.wikipedia.org/wiki/Apple
FAROO WIKIPEDIA
Apple Trees are Cultivated Worldwide

2 **Apple reports record iPhone revenue over the holidays while ...**
Apple Inc. is an American multinational technology company headquartered in Cupertino, California, that designs, develops, and sells consumer electronics, ... Number of employees: 137,000 (2019) Number of locations: 500+ retail stores (2019) ...
https://www.gsmarena.com/apple_reports_
GOOGLE
Apple Store iPhone Revenue Apple Reports Record

Fig. 1. User interface of commercial SRC system – CARROTSEARCH

SRC techniques have been developed for many years not only in academic community but also in commercial systems (Figure 1). However, SRC algorithms are far from being perfect because relatedness of information changes over time. Daily events happen and change facts as well as connection between information. For example, relatedness of terms "Donald Trump" and "US President" might be increased after election in 2016. Now, search results containing phrase "US President", are more likely grouped with search results containing phrase "Donald Trump".

To effectively cluster search results, many researchers have applied external knowledge resources and Wikipedia is one of them. In this paper, Wikipedia is used to extract topics from title and snippet of document. Affinity propagation clustering algorithm is used to cluster search results instead of standard k-mean algorithm.

The rest of paper is organized as follows: Section 2 describes the related works. Section 3 explains proposed method in details. Some experiments and results are discussed in the last section.

2. Related work

Recently, many researchers have focused on the problem of Web Results Clustering. A detailed survey of various SRC algorithms has been presented in [5]. These algorithms are mainly divided based on the ability of labeling clusters from data-centric algorithms to description-centric algorithms. A common step to all SRC system is input preprocessing. Its aim is to convert contents of search results into a suitable form for clustering algorithms. Usually, search results go through preprocessing steps like: tokenization, stemming and eventually enrich document representation.

For most of traditional text based clustering methods like SuffixTree clustering (STC), search results are represented as "bag of words". This method takes into account only shared phrases between different documents and ignores the semantic relationships between key terms. Two search results on same topic having synonyms or semantically associated terms might be assigned to different clusters. This problem was partially solved by using ontology like WordNet as background knowledge [9].

In WordNet, synonyms are grouped into synsets and all synsets are connected to others by semantic relations. By using these synsets, LINGO algorithm [8] represents search results as "bag of synsets" before clustering using a modified version of k-mean algorithm. A disadvantage of using WordNet is that its word-sense disambiguation (WSD) ability is not effective and manual creation of knowledge resource like WordNet is expensive and time consuming. Recently, an automated creation WSD framework called Babelfy has gained more attention since its core concepts and relations are collected from both WordNet and Wikipedia. But when applied in SRC problem, it seems that quality improvement is not significant, as reported in.

Using Wikipedia to extract concepts from text, authors in [7] consider document titles as concepts and match them with search results. Terms in search results also were used to create a graph of candidate clusters. A search result is attached to a candidate cluster if it contains specific term or concept representing for this cluster. Then this graph is clustered by finding its subgraphs.

Inspired by this work, Tagme [6], a state of the art topic annotator for short texts is utilized in this paper to process search results before being clustered by affinity propagation algorithm – a well-known algorithm for small dataset.

3. Proposed method

SRC is post-retrieval clustering process that clusters search results based on their meaning. Usually, input for SRC system is a list of short texts from several first pages returned by search engine. Therefore, two characteristics of this problem are: small dataset and ambiguity of short texts.

Proposed algorithm contains two main stages: topics extraction and clustering. In the first stage, the main idea is describing a search result by a *bag of topics*. To get this done, document's title and snippet returned by search engine are annotated by Tagme system. Not only snippet but also title is used to extract topic from document. Because title of document often contains its main topic while snippet is only a small fragment of document (two or three lines), so it is worth to include topics extracted from title to representation of document. After this step, each search result is represented as a set of pairs: ($topic^{ij}$, $score^{ij}$) where $score^{ij}$ is a real number from 0 to 1 and stands for probability of i-th document is about $topic^{ij}$.

In second stage, affinity propagation clustering algorithm is employed. Main advantage of this algorithm is that it does not require beforehand setup meta parameters like others data-centric clustering algorithm. For example, number of clusters in k-mean algorithm. Instead, this algorithm creates clusters by sending message between pair of samples until convergence. The message sent between pairs represent suitability for one sample to be the exemplar of the other. The algorithm takes as input a real matrix S where $s\,(i, j)$ determines similarity of i-th and j-th document, $s\,(i, i)$ is ability of choosing i-th document as exemplar of cluster. A disadvantage of this algorithm is time complexity and memory complexity, but this is quite acceptable for SRC problem when number of search results to cluster is not too much.

To compute similarity of documents, we use topics annotated in first stage. Assume that, i-th and j-th document are represented by set $\{(topic^{ip}, score^{ip})\}$, $\{(topic^{jq}, score^{jq})\}$ respectively, similarity between i-th and j-th documents is defined as follow:

$$s(i,j) = \\ = \sum score^{ip} * relatedness(\,topic^{ip}, topic^{jq}) * \\ * \, score^{jq} \qquad (1)$$

where $relatedness\,(topic^{ip}, topic^{jq})$ is similarity between two Wikipedia topic. This measurement is also taken from Tagme system. After clustering, each cluster is assigned a label by topic which has maximum score among all topics in this cluster.

4. Experiment and conclusion

Experiment is conducted on AMBIENT (AMBIguous ENTries) dataset. The dataset contains 44 ambiguous queries; the average length of queries is 1.27. Each query contains 100 search results, and golden standard clustering (created by human) [10, 11].

Clustering evaluation is a difficult issue but four following measures are widely accepted: Rand Index (RI), Adjusted Rand Index (ARI), Jaccard Index (JI) and F1 measure. Evaluator is taken from Evaluating Word Sense Induction & Disambiguation within An End-User Application contest (SemEval2013-Task11).

The proposed method is compared in terms of clustering quality with the known Lingo method. The quality of clustering is tested based on four typical measures.

Tab. 1. The proposed method is compared in terms of clustering quality with the known Lingo method. The quality of clustering is tested based on four typical measures

Method	**RI**	**ARI**	**JI**	**F1**
Lingo	62,52	**18,09**	**30,76**	49,01
Proposed method	**63,39**	16,64	25,46	**55,26**

Python package of Tagme system [12] is used for extraction phrase and *sklearn* implementation [13] of affinity propagation clustering algorithm is used to cluster documents. Proposed method overcomes algorithm Lingo in two criteria but not in ARI and JI measurement. The improvement is also not too much: 1% for RI and about 5% for F1. For ARI criteria, 0 is putting data point in random clusters, both Lingo and proposed method do not exceed threshold 0.2 which mean they are quite close to randomness. Low value for ARI criteria confirm that it is very difficult to recreate clusters like human do. There is also problem with performance when use Tagme system – query external server is time consuming.

Comparing performance of methods for specific problem is not a trivial task, especially when several metrics are used simultaneously. For example, in this paper there are four measures: RI, ARI, JI and F1. In this situation, the global quality of the considered methods can be determined (and compared) using multi-criteria analysis methods [3]. If these measures are considered as axes in a decision space, so each method is one decision point. In multi-criteria analysis, an ideal point is calculated from decision points, its coordinates are maximum value along each axis. A method is considered better than others if its decision point has shorter distance from ideal point [2, 3]. By using multi-criteria analysis in this space, comparing methods is more objective. For Lingo and proposed method, their corresponding decision points are:

LM = (62.52, 18.09, 30.76, 49.01)

– Lingo method

PM = (63.39, 16.64, 25.46, 55.26)

– Proposed method

*Y** = (63.39, 18.09, 30.76, 55.26)

– Ideal point (as an utopia – ideal method [2, 3])

Distances (p – distances) from ideal point can be obtained by Minkowski distance for $p \geq 1$ [2, 3] as a $\left\| Y^{*} - LM \right\|_{p}$ and $\left\| Y^{*} - PM \right\|_{p}$.

We can treat Minkowski's distance (after normalization [3]) as the *p*-similarity of the analyzed methods to the ideal method. The smaller the distance to ideal point is – the better the analyzed method is. We use most often $p = 1, 2, \infty$, so we have:

for $p = 1$:

$\left\| Y^* - LM \right\|_p = 7.12$ and $\left\| Y^* - PM \right\|_p = 6.75$,

for $p = 2$:

$\left\| Y^* - LM \right\|_p = 6.31$ and $\left\| Y^* - PM \right\|_p = 5.49$,

for $p = \infty$:

$\left\| Y^* - LM \right\|_p = 6.25$ and $\left\| Y^* - PM \right\|_p = 5.30$.

In all three case:

$$\left\| Y^* - LM \right\|_p \quad > \quad \left\| Y^* - PM \right\|_p .$$

It means: *PM* is better than *LM*. From this results, proposed method can be considered better than Lingo method in global multi-criteria assessment.

In this paper, Wikipedia is used as external knowledge resource to represent short texts as “bag of topics” before applying affinity propagation clustering algorithm. Although proposed method does not significantly overcome the state of the art – Lingo algorithm but its performance is comparable with it. Testing similar clustering tasks with the presented method gives analogous results. Therefore, the above concept of improving clustering seems to be promising. In the future, we plan to use other document similarity scoring schema [1, 3, 4] to see if it can help to improve clustering quality.

Acknowledgement: This work is supported by Military University of Technology, Young Scientists Development Program (RMN 864).

5. Bibliography

[1] Ameljańczyk A., “Teoretyczne aspekty badania podobieństwa obiektów w problematyce rozpoznania wzorców”, in: *Problemy modelowania i projektowania opartych na wiedzy systemów informatycznych na potrzeby bezpieczeństwa narodowego*, T. Nowicki, Z. Tarapata (Eds.), pp. 9–22, WAT, Warszawa 2014.

[2] Ameljańczyk A., “Multicriteria similarity models for medical diagnostic support algorithms”, *Bio-Algorithms and Med-Systems*, Vol. 9, 1–7 (2013).

[3] Ameljańczyk A., *Multicriteria optimization in control and management problems*, Zakład Narodowy im. Ossolińskich, 1984.

[4] Brendan J., Frey B.J. and Dueck D., “Clustering by Passing Messages Between Data Points”, *Science*, Vol. 315, 972–976 (2007).

[5] Carpineto C., Osinski S., Romano G., Weiss D., “A Survey of Web Clustering Engines”, *ACM Computing Surveys*, Vol. 41, No. 3, Art. 17 (2009).

[6] Ferragina P., Scaiella U., “Fast and Accurate Annotation of Short Texts with Wikipedia Pages”, *IEEE Software*, Vol. 29(1), 70–75 (2012).

[7] Jinarat S., Haruechaiyasak Ch., Rungsawang A., “Graph-Based Concept Clustering for Web Search Results”, *International Journal of Electrical and Computer Engineering* (*IJECE*), Vol. 5, No. 6, 1536–1544 (2015).

[8] Osiński S. and Weiss D., “A Concept-Driven Algorithm for Clustering Search Results”, in: *IEEE Intelligent Systems*, Vol. 20, Issue 3, 48–54, IEEE 2005.

[9] Sameh A., Kadray A., “Semantic Web Search Results Clustering Using Lingo and WordNet”, *International Journal of Research and Reviews in Computer Science* (*IJRRCS*), Vol. 1, No. 2, 71–76 (2010).

[10] Carpineto C., Romano G., *Ambient dataset*, http://search.fub.it/ambient/.

[11] Evaluating Word Sense Induction & Disambiguation within An End-User Application: https://www.cs.york.ac.uk/semeval-2013/task11/.

[12] https://github.com/marcocor/tagme-python.

[13] https://scikit-learn.org/.

[14] https://en.wikipedia.org/wiki/Web_search_query.

Klasteryzacja wyników wyszukiwania z wykorzystaniem Wikipedii

C. TRAN, A. AMELJAŃCZYK

W pracy przedstawiono propozycję nowej metody klasteryzacji wyników wyszukiwania. Metoda wykorzystuje zewnętrzny zasób wiedzy, którym jest Wikipedia. Wikipedia – największa encyklopedia – to darmowy i popularny zasób wiedzy służący do wydobywania tematów z krótkich tekstów. Podobieństwa między dokumentami są obliczone na podstawie podobieństwa między tymi tematami. Następnie algorytm klasteryzacji, bazując na propagacji powinowactwa, jest wykorzystywany do grupowania wyników wyszukiwania w Internecie. Proponowana metoda jest testowana przez zbiór danych AMBIENT i oceniana w ramach eksperymentalnych narzędzi dostarczonych przez konkurs SemEval-2013. W artykule zaproponowano również nową metodę porównywania globalnej wydajności algorytmów z wykorzystaniem analizy wielokryterialnej.

Słowa kluczowe: propagacja powinowactwa, podobieństwo dokumentów, ocena wielokryterialna, wyniki wyszukiwania w sieci, Wikipedia.