Date of acceptance Grade

Instructor

Advances in news event aggregation methodology

Phuong Nguyen

Helsinki October 17, 2017

UNIVERSITY OF HELSINKI
Department of Computer Science

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Laitos — Institution — Department
Faculty of Science	Department of Computer Science

Tekijä — Författare — Author
Phuong Nguyen

Työn nimi — Arbetets titel — Title

Advances in news event aggregation methodology

Oppiaine — Läroämne — Subject
Computer Science

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
	October 17, 2017	6 pages + 0 appendices

Tiivistelmä — Referat — Abstract

News event grouping methods have been developed constantly over the year to aid news comsumption, especially in the age of information. This report describes the progress and development in response to the modern context of digital information explosion.

Avainsanat — Nyckelord — Keywords
news event, aggregation, named entity, keyword

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — övriga uppgifter — Additional information

Contents

1 Introduction

During the vigorous growth of information in the 21st century, we have witnessed many challenges that require our attention. Questions have risen concerning factual accuracy, relevance, privacy, etc. While the abundance of information is useful in many cases to human society, the excess is very often the seed that create struggle in finding beneficial information that satisfy our demands. This happens especially in the field of news reporting, where the evolution of news delivering has gone far from the traditional physical newspaper to embrace the new mainstream called digital news.

To illustrate the challenges we are facing, imagine what happens every time a new event occurs, regardless of where in the world. There is almost always a target group of people who need access to the story regarding the event. The story needs to be classified in accordance to target's interests. However, some events might generate a significant amount of interest, which in turn causing a huge corresponding amount of different stories that refer to the same event. This inevitably leads to the problem where audiences have to deal with a mass amount of similar information, while the truthfulness of the stories related to the event become questionable. However, to properly dispatch news in a timely manner to the right audience, with first-class correctness and without redundancy, various strategies have been developed over the time.

In order to understand how the aforementioned problems are being solved, basic methods and several works in the field will be reviewed as a brief introduction to the development of news aggregation methodology. The report will conclude with a summary of the general work that have been done so far, contemporary limitations and a brief discussion on the prospect in the coming years.

2 News clustering

Clustering news event is a specific problem resided within a bigger topic, text clustering. Text clustering is a method of grouping similar textual data into categories that can be used for effective information retrieval. The field has been studied extensively to mine textual data and thus many different methods have been invented.

Regardless of the method used, in order to perform a clustering method on a set of documents, each document should be converted into a processable representation. A very first basic step to produce this representation is to remove stop words and then apply stemming to remaining words. A vectorization method is then used to facilitate distance calculation between documents, after which, specialized algorithms are employed to categorize and retrieve documents.

2.1 Trimming and Stemming

Each language has its own redundant and frequent words that generally do not contribute to the distinctiveness of any arbitrarily written topic. In English, words such as 'a', 'the', etc. are prime examples. However, there are stop words that can appear infrequently as well, with 'nevertheless' standing as a good example. Removing these stop words helps reducing the number of processing related to them, while placing more focus on recognizing the unique features of a document [RU11, p. 8].

Stemming is a method of reducing or transforming a word back into its base form, so that all words that derived from the same root can be expressed using one shared stem. This is done to ease computational linguistic problems as mathematical analysis of a text usually requires matched stems [Lov68, p. 22]. By applying a stemming algorithm, words that shared the same origin are then treated the same way, thus aiding frequency analysis methods such as TF-IDF.

2.2 TF-IDF

Two different documents cannot be directly compared to each other computationally. They need to be converted into mathematical representations that support distance calculation. A very common approach to achieve this is by vectorizing documents using TF-IDF (term frequency-inverse document frequency) weighting

scheme [BGLB16, p. 319].

In this weighting scheme, each word is counted for its number of occurrence so its frequency can be calculated. This corresponds to the TF (term frequency) part of the scheme. By inverting the frequency of each word, the result is a normalized document with the most occurring words being the least important. This is the IDF (inverse document frequency) part of the algorithm, which promote a document's uniqueness based on its least common occurring words [AZ12, p. 80]. Due to this nature, TF-IDF is very often used to create a distance between a document and others, as well as ranking them in order of relevance to a specific topic.

2.3 Contextual analysis

Traditionally, by using a simple keyword to represent a topic, relevant documents or information can be retrieved. However, as suggested by Lam et al., the use of concept terms and named entities together with the classic story keywords can form a context which demonstrates better performance than the traditional implementation where only story keywords are employed [LMWY01, p. 525].

With a simple news system where only keyword is used, the system displays struggle in classifying two stories representing two different events, which contain similar keywords but having different styles and words. This is a intrinsic nature of human language where vocabulary can be substituted, thus creating barriers to achieve good performance in news clustering system [LMWY01, p. 527].

To improve the performance, named entities, which represent person, organization and person name, etc. were coupled with classical story keywords and concept terms. Concept terms are statistical analysis results of a story, extracted by utilizing a concept database according to the connection between stories in the database and sentences contained within the story. The results show a good performance with small chance of false alarm, albeit a higher stories missing probability in certain cases [LMWY01, p. 544–545].

2.4 Salience of named entities

Another research group, led by Roman Yangarber, demonstrates that by using 'salience' feature, which is a measure of named entities' importance, can yield substantial performance gain compared to the standard story keyword feature. The

group also concludes that named entities alone are better features than words [YEP+17, p. 1096].

For named entities, PULS (an on-line information extraction system) is used to extract them from stories, using more than a thousand patterns. Each named entity is then assigned a type as it is considered useful to utilize early classification within a set of clustered documents for performance advantage. Named entities are also mentioned with different frequencies within an article. Most important entity are usually stated in the beginning, especially in the title and description of an article, followed by more elaboration within the first few sentences. Less important and less relevant entities are cited later on with generally lower frequencies [YEP+17, p. 1098].

The results show improved performance when it comes to detection of related stories from many different news articles. This gain is not obtained by using 'salience' feature alone, but also adjustment on clustering method, which combine different vectorized representations to progress beyond the base method. Nevertheless, user metrics can also be explored along with other clustering algorithms to see the changes in the system, while possibly further improve the accuracy [YEP+17, p. 1104].

3 Conclusion

The work in text clustering field or specifically news clustering has progressed significantly over the years. Research groups have gone from using conventional simple keywords as unique features to making use of various contextual settings and semantic connections to determine the relation between documents. These advances have made it easy for general information retrieving and became particularly useful in news distributing systems where a number of constraint have to be met.

As far as accuracy is concerned, grouping precision can still be improved despite the advancement in the field. The hidden context of textual data can also be further looked into, as news data, for example, can have various implicit features that are yet to be found, compared to a standard text. The mathematical methods used to represent and cluster data are also something worth investigating. The research in the field, therefore, remains opened up for many possibilities.

References

AZ12 Aggarwal, C. C. and Zhai, C. *A Survey of Text Clustering Algorithms*, pages 77–128. Springer US, Boston, MA, 2012. URL https://doi.org/10.1007/978-1-4614-3223-4_4.

BGLB16 Beel, J., Gipp, B., Langer, S. and Breitinger, C., Research-paper recommender systems: a literature survey. *International Journal on Digital Libraries*, 17,4(2016), pages 305–338. URL https://doi.org/10.1007/s00799-015-0156-0.

LMWY01 Lam, W., Meng, H. M. L., Wong, K. L. and Yen, J. C. H., Using contextual analysis for news event detection. *International Journal of Intelligent Systems*, 16,4(2001), pages 525–546. URL http://dx.doi.org/10.1002/int.1022.

Lov68 Lovins, J. B., Development of a stemming algorithm. *Mech. Translat. & Comp. Linguistics*, 11,1-2(1968), pages 22–31.

RU11 Rajaraman, A. and Ullman, J. D. *Data Mining*, pages 1–17. Cambridge University Press, 2011.

YEP+17 Yangarber, R., Escoter, L., Pivovarova, L., Du, M. and Katinskaia, A., Grouping business news stories based on salience of named entities. *EACL*, 2017, pages 1096–1106.