Named-Entity-based Linking and Exploration of News using an Adapted Jaccard Metric

Tom De Nies, Jasper Verplanken, Ruben Verborg, Wesley De Neve, Erik Mannens, and Rik Van de Walle

In: Proceedings of the Workshop on Negative or Inconclusive Results in Semantic Web (NoISE2015), 2015.

http://ceur-ws.org/Vol-1435/NoISE2015_paper_2.pdf

De Nies, T., Verplanken, J., Verborg, R., De Neve, W., Mannens, E., and Van de Walle, R. (2015). Named-Entity-based Linking and Exploration of News using an Adapted Jaccard Metric. ***Proceedings of the Workshop on Negative or Inconclusive Results in Semantic Web (NoISE2015)***

Named-Entity-based Linking and Exploration of News using an Adapted Jaccard Metric

Tom De Nies[1], Jasper Verplanken[1], Ruben Verborgh[1], Wesley De Neve[1,2],
Erik Mannens[1], and Rik Van de Walle[1]

[1] {firstname.lastname}@ugent.be
Ghent University – iMinds – Multimedia Lab, Belgium
[2] KAIST – IVY Lab, Republic of Korea

Abstract. In this paper, we propose a semantically enabled news exploration method to aid journalists in overcoming the information overload in today's news streams. To achieve this, our approach semantically tags news articles, calculates their relatedness through their similarity based on these tags, and creates an article graph to be browsed by an end-user. Based on related work, the Jaccard metric seemed very suitable for this task. However, when we evaluated this similarity measure through crowdsourcing on a set of 120 article pairs, the results were only acceptable in the lower levels of relatedness, with unpredictable errors elsewhere. This reveals a need for better ground-truth data, and calls for clarification of the semantics of relatedness and similarity, and their relation.

1 Introduction

Nowadays, there is an abundance of information that journalists of various news organizations have to process. Because painful errors can be made when valuable news information is skipped, journalists are often forced to iterate over news items sequentially. Combined with the increase in data channels, data volume and demand for 24/7 news delivery, this puts a significant pressure on journalists. Therefore, the need arises for more intelligent navigation techniques.

In this paper, we propose a browsing method for news exploration based on semantic relatedness. The main research question we aim to address with this approach is: *is it possible to capture relatedness using semantic similarity?*. The hypothesis we test in this paper is that *it is possible to capture relatedness using an adaptation of the Jaccard metric.* To evaluate this, we *tag* news articles with named entities and measure their similarity using these entities. That way, links are formed between related articles, and a network of news is created for the journalist to browse through.

1.1 Related Work

The Jaccard metric has been successfully adapted for such a linking scenario before. For example, in [2], it is adapted to discover meaningful connections between different concepts in Linked Data. Similarly, it was evaluated in an entity linking scenario by Ceccarelli et al. [1], and in a recommendation scenario by Passant [3]. In all these cases, the Jaccard metric showed promising results, especially considering its ease of calculation and applicability in various scenarios. Therefore, we deemed it appropriate for our proposed approach. However, our evaluation on an article-set from The Guardian (see Section 3) reveals unpredictable errors, with the only acceptable results in the lower levels of relatedness.

2 Proposed Approach

In this section, we will present a solution that aims at coping with the problem of information overload in a professional journalism environment. To achieve this, we argue that the listing of newsworthy items is no longer sufficient to visualize the detailed connections that may exist between various media items. A *graph* would correspond more accurately to the complex associations made within a human mind during the search process. Therefore, we propose a weighted graph-based view, as illustrated by Figure 1. The nodes are media items, either textual or visual, and the links carry attributes and weights. Our approach can be summarized in three steps: 1) semantic tagging, 2) measuring similarity, and 3) visualization.

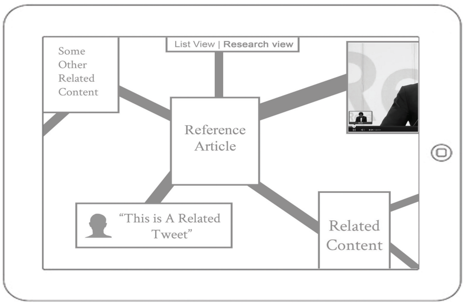

Some Other Related Content
List View | Research view
Reference Article
"This is A Related Tweet"
Related Content

Fig. 1. Conceptual illustration of a data-exploration browser with network-based navigation and support for multiple media types.

2.1 Semantic Tagging

Named entities (NEs) are named semantic concepts that appear in a piece of text. They identify keywords that carry a semantic payload, which unambiguously describes what the keyword means. For example, in the sentence *"I saw George in Washington today"*, the entity 'George' refers to a *person*, and 'Washington' to a *location* (e.g., the state or city). However, in the sentence *"I saw George Washington today"*, the entity 'George Washington' refers to a person (the first president of the US). Resolving this kind of ambiguity is the main subject of many research efforts in named entity recognition and natural language processing. Recent evaluations such as Van Erp et al. [5] and GERBIL [4] provide an insight into the current state-of-the-art in that field. In this paper, however, we focus on the *usage* of named entities, rather than on their *recognition*.

2.2 Measuring Similarity

The metric used to assign weights to the edges can be any similarity measurement involving the properties of both news items. We argue that the collection of all NEs found in an article carries a great semantic payload that can be used for this purpose. More specifically, we use an adaptation of the Jaccard metric, with the NEs recognized in a document as its features. We calculate the NE-based Jaccard similarity between two articles a and b as follows:

$$Jaccard_{NE}(a,b) = \frac{|N(a) \cap N(b)|}{|N(a) \cup N(b)|}, \tag{1}$$

where $N(x)$ is the set of all entities recognized in document x.

2.3 Visualization

In order to provide a more *immersive* navigation experience to the user, we chose a visualization optimized for touch screens for our proof-of-concept implementation. When starting the news exploration application, a user is presented with an arbitrary reference article, linked to four suggested articles, as illustrated in Figure 2. The suggested articles are selected by calculating the similarity of the reference article to the other articles in the dataset, as described in Section 2.2, and selecting the four with the highest similarity. When the user navigates to one of these four suggested articles, this article is put in place of the reference article, and the process is repeated. That way, the user can effectively explore the dataset and discover new articles.

3 Evaluation

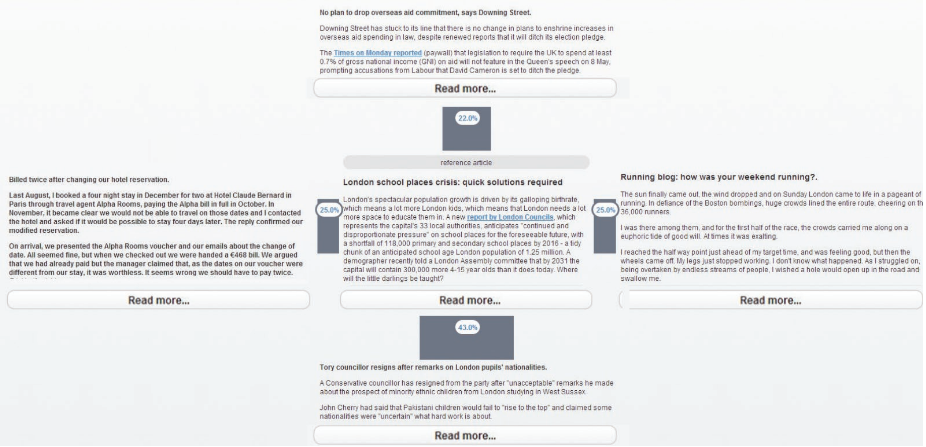
To evaluate our approach, we gathered a set of 851 English news articles[3], over the course of one week from the online newspaper The Guardian. All articles were semantically tagged with NEs[4] using the NER service AlchemyAPI[5]. From this set, we randomly selected 30 articles, which we used as reference articles. For each of these 30 articles, we then used our approach with the NE-based Jaccard similarity to find four relevant articles in the dataset. In other words, we generated a set of 120 links in total (4 for each of the 30 articles).

We then performed an evaluation using the Amazon Mechanical Turk (AMT) platform. In total, 120 Human Intelligence Tasks (HITs) were created, one for each linked article pair. In each HIT, the AMT worker was presented an article-pair, and was asked to rate the relatedness of the articles' content on a 5-point Likert scale. The scale had the following scores:

[3] List of URIs of the article corpus used: `http://bit.ly/1hqdKi5`

[4] URIs + extracted NEs: `http://bit.ly/1bK6CoE`

[5] `http://www.alchemyapi.com/`



No plan to drop overseas aid commitment, says Downing Street.

Downing Street has stuck to its line that there is no change in plans to enshrine increases in overseas aid spending in law, despite renewed reports that it will ditch its election pledge.

The Times on Monday reported (paywall) that legislation to require the UK to spend at least 0.7% of gross national income (GNI) on aid will not feature in the Queen's speech on 8 May, prompting accusations from Labour that David Cameron is set to ditch the pledge.

Billed twice after changing our hotel reservation.

Last August, I booked a four night stay in December for two at Hotel Claude Bernard in Paris through travel agent Alpha Rooms, paying the Alpha bill in full in October. In November, it became clear we would not be able to travel on those dates and I contacted the hotel and asked if it would be possible to stay four days later. The reply confirmed our modified reservation.

On arrival, we presented the Alpha Rooms voucher and our emails about the change of date. All seemed fine, but when we checked out we were handed a €468 bill. We argued that we had already paid but the manager claimed that, as the dates on our voucher were different from our stay, it was worthless. It seems wrong we should have to pay twice.

London school places crisis: quick solutions required

London's spectacular population growth is driven by its galloping birthrate, which means a lot more London kids, which means that London needs a lot more space to educate them in. A new report by London Councils, which represents the capital's 33 local authorities, anticipates "continued and disproportionate pressure" on school places for the foreseeable future, with a shortfall of 118,000 primary and secondary school places by 2016 - a tidy chunk of an anticipated school age London population of 1.25 million. A demographer recently told a London Assembly committee that by 2031 the capital will contain 300,000 more 4-15 year olds than it does today. Where will the little darlings be taught?

Running blog: how was your weekend running?.

The sun finally came out, the wind dropped and on Sunday London came to life in a pageant of running. In defiance of the Boston bombings, huge crowds lined the entire route, cheering on th 36,000 runners.

I was there among them, and for the first half of the race, the crowds carried me along on a euphoric tide of good will. At times it was exalting.

I reached the half way point just ahead of my target time, and was feeling good, but then the wheels came off. My legs just stopped working. I don't know what happened. As I struggled on, being overtaken by endless streams of people, I wished a hole would open up in the road and swallow me.

Tory councillor resigns after remarks on London pupils' nationalities.

A Conservative councillor has resigned from the party after "unacceptable" remarks he made about the prospect of minority ethnic children from London studying in West Sussex.

John Cherry had said that Pakistani children would fail to "rise to the top" and claimed some nationalities were "uncertain" what hard work is about.

Fig. 2. The news exploration screen, showing a reference article surrounded by four article suggestions. The similarity score is indicated by the thickness of the links.

1	2	3	4	5
Not related at all	Slightly related	Moderately related	Highly related	Completely related (almost the same)

As a preventive measure against spam, we also asked the AMT workers to explain *why* they thought the articles were (un)related, as well as a short summary of the reference article. Additionally, we filtered out all HITs that were submitted in less than 30 seconds, had an empty explanation or summary, or exhibited an obvious indication of being automatically generated (such as identical, generic responses over multiple HITs). Each article pair was evaluated by 10 different users, leading to 1200 evaluations in total. However, after applying the spam-control measures, we dismissed 76 answers, resulting in a final set of 1124 evaluations.

When normalized between 0 and 1, the evaluations on the Likert scale allow us to quantitatively measure the difference between the human assessment of relatedness, and the automatic assessment of similarity using our approach. We define the average similarity score of all article pairs as $S_{Jaccard}$, and the average evaluation score S_{Likert} as follows:

$$S_{Likert} = \frac{\sum_{e \in E}(e-1)}{|E| \times 4}. \tag{2}$$

Here, E is the set of HIT evaluations for an article pair, and $e \in E$ one of those evaluations, its value ranging from 1 to 5. This means that evaluations of Likert level 5, 4, 3, 2 and 1 will correspond to scores of 1, 0.75, 0.5, 0.25 and 0, respectively. We observed an average absolute difference $|S_{Likert} - S_{Jaccard}|$ of *0.198* between all the evaluations per article pair to the assessments made by our approach, which corresponds to a difference of 1 Likert level at most. However, we also observed that the error varied positively or negatively for each article pair, meaning that it cannot be automatically corrected for.

Apart from this value, we also calculated *precision* and *recall* values for each level of the Likert scale, by observing the number of **true positives (TP)**, **false positives (FP)**, and **false negatives (FN)** per Likert level, defined as follows:

TP: number of article pairs correctly assigned to this Likert level;
FP: number of article pairs incorrectly assigned to this Likert level;
FN: number of article pairs incorrectly assigned to a different Likert level.

We assign each Likert level to a range of possible values, as indicated in Table 1. The precision (P) and recall (R) of each Likert level can now be calculated as $P = \frac{TP}{TP+FP}$ and $R = \frac{TP}{TP+FN}$, respectively.

We calculated these values for each Likert level, as shown in Table 1. When observing these results, it is clear that the precision and recall of the Jaccard metric is only acceptable in the lowest range of relatedness as assessed by the users. In fact, it seems that the majority of article pairs were classified by the approach in the $[0, 0.4[$ range. This is surprising, because the dataset consisted of the articles deemed most relevant to the reference articles by the approach. However, the average score assigned to all article pairs by the AMT workers was as low as 0.237, which corresponds to a Likert level of "2: slightly related". This means that our dataset was biased towards less related articles, and that the approach simply did not have enough highly related articles to choose from.

Another possible explanation for the lower precision in the ranges above 0.2, is that the NE-based Jaccard measure does not scale in the same way as the human assessment. Although the average absolute difference of nearly *20%* between the human and the automatic assessment potentially supports this, further experiments will need to be performed in future work.

Lastly, we observed a small correlation between the minimum number of NEs recognized in the article pairs and the absolute difference in measured relatedness score and calculated similarity score. This indicates that the approach might not be suitable for texts where few or no NEs can be detected. Additionally, this stresses the importance of the quality of the NER service.

Table 1. Precision (P) and recall (R) values for each Likert level, mapped to its corresponding range of assessment scores. Additionally, the no. of human assessments (HA), automatic assessments (AA), true positives (TP), false positives (FP), and false negatives (FN) is shown. The only acceptable P&R values are those for Likert level 1.

Likert level	Score Range	HA	AA	TP	FP	FN	**P**	**R**
1	$[0, 0.2[$	73	61	38	23	35	**0.623**	**0.521**
2	$[0.2, 0.4[$	18	50	7	43	11	0.140	0.389
3	$[0.4, 0.6[$	18	8	1	7	17	0.125	0.056
4	$[0.6, 0.8[$	9	1	1	0	8	1.0	0.111
5	$[0.8, 1]$	2	0	0	0	2	0	0

4 Discussion and Future Work

The results in Section 3 indicate that the NE-based Jaccard metric is neither completely suitable nor entirely unsuitable for our approach. An average difference of one Likert level with the human assessment of relatedness is not catastrophic, but its unpredictability in the positive or negative sense makes it impossible to automatically correct for. The metric does show promise in the range of the lower relatedness scores, but it remains unclear whether this was due to the bias of the dataset. Therefore, we must call these results inconclusive.

An important lesson learned from this research is that an important distinction is to be made between *similarity* and *relatedness*. Like many recommendation approaches, we assumed that the two show significant correlation. However, this is not always guaranteed, especially in the case of news. An article-pair can be less similar, yet very related. For example, an article about the investigation of a United Airlines crash, and an article about the investigation of a Turkish Airlines crash might be considered similar, but arguably not so related. Analogously, an article about the investigation of a airplane crash and an article about safety measures in aviation might be considered very related, yet not so similar.

In future work, we aim to investigate this distinction more elaborately by clearly defining the semantics of relatedness, and its relation with similarity. An additional challenge will be to make this distinction clear to humans using the platform, and correctly scale the automatic scoring to match the human interpretation. The influence of the NER quality and document length must also be looked into. Finally, the usability of our news exploration approach remains untested, and must be investigated before the application can be considered for practical use in a real-world scenario.

Acknowledgments. The research activities in this paper were funded by Ghent University, iMinds (by the Flemish Government), IWT Flanders, FWO-Flanders, the European Union, and the Belgian Industrial R&D MSc thesis programme.

References

[1] Ceccarelli, D., Lucchese, C., Orlando, S., Perego, R., Trani, S.: Learning relatedness measures for entity linking. In: Proc. of the 22nd ACM international conference on Conference on information & knowledge management. pp. 139–148. ACM (2013)

[2] De Vocht, L., Coppens, S., Verborgh, R., Vander Sande, M., Mannens, E., Van de Walle, R.: Discovering meaningful connections between resources in the web of data. In: Proceedings of the 6th Workshop on Linked Data on the Web (LDOW) (2013)

[3] Passant, A.: Measuring semantic distance on linking data and using it for resources recommendations. In: AAAI Spring Symposium: Linked Data Meets Artificial Intelligence. vol. 77, p. 123 (2010)

[4] Usbeck, R., et al.: GERBIL – general entity annotation benchmark framework. In: 24th WWW conference (2015)

[5] Van Erp, M., Rizzo, G., Troncy, R.: Learning with the Web: Spotting Named Entities on the Intersection of NERD and Machine Learning. In: 3rd workshop on Making Sense of Microposts. pp. 27–30 (2013)