A News Chain Evaluation Methodology along with a Lattice-based Approach for News Chain Construction

Mustafa Toprak
mustafatoprak@iyte.edu.tr
Izmir Institute of Technology,
35430, Urla, Izmir, Turkey

Özer Özkahraman
ozerozkahraman@outlook.com
KTH Royal Institute of Technology,
Stockholm 100-44, Sweden

Selma Tekir
selmatekir@iyte.edu.tr
Izmir Institute of Technology,
35430, Urla, Izmir, Turkey

Abstract

Chain construction is an important requirement for understanding news and establishing the context. A news chain can be defined as a coherent set of articles that explains an event or a story. There's a lack of well-established methods in this area.

In this work, we propose a methodology to evaluate the "goodness" of a given news chain and implement a concept lattice-based news chain construction method by Hossain et al.. The methodology part is vital as it directly affects the growth of research in this area. Our proposed methodology consists of collected news chains from different studies and two "goodness" metrics, $minedge$ and $dispersion\ coefficient$ respectively. We assess the utility of the lattice-based news chain construction method by our proposed methodology.

1 Introduction

A news story is a compelling organization of news to give an overall idea about an event or a set of related events. Generally this organization follows a time order and has topical coherence. The most common approach to construct a news chain is called "connecting the dots" (Shahaf and Guestrin, 2010). In this approach, there are predetermined start and end points and the task is to find a coherent sequence of articles between them.

In today's substantial news flow, tracking all news to understand an event or establish connections between related events is a challenge. Thus, automated mechanisms are needed to construct news chains and to support users in making news stories.

Our intended contribution is thus twofold: First, there's a need for a methodology in order to assess the quality of given news chains. Second, we implemented a state-of-the-art method that is based on the concept lattice representation of the news articles and evaluated its effectiveness in a more extensive way than the provided and additionally using our methodology.

In order to establish a news chain assessment methodology, we refer to two independent "goodness" metrics that are proposed, and experimentally validate and compare them in the same experimental setup. As far as we know, there is no such unifying experimental design that has a set of news chains and run of these metrics under the same conditions. Thus, we provide an evaluation regarding the utility of the proposed metrics in the quality assessment of news chains. Our finding is that **minedge** metric proposed by (Shahaf and Guestrin, 2010) behaves in a consistent way, but **dispersion coefficient** metric suggested by (Hossain et al., 2011) does not serve the purpose as expected.

As for the task of news chain construction, utilizing concept lattice-based representation of news articles (Hossain et al., 2011) is in accordance with our intuition. When we considered order relations, the sequence of articles that form a chain has a linear order. This linearity is provided by a total order relation. As partial order relations are more generic than their total order counterparts, our idea was to define a partial order relation over the set of articles and obtain a pool of news chain candidates out of the generated hierarchy. Thus, we create partially ordered news articles using their content. In this sense, we use a proposed pruning and heuristic (Hossain et al., 2011) to extract useful news chains out of the candidate pool. We evaluate the "goodness" of the constructed chains by the use of established methodology.

Proceedings of the 2017 EMNLP Workshop on Natural Language Processing meets Journalism, pages 95–99
Copenhagen, Denmark, September 7, 2017.

2 Related Work

News chain construction aims to discover hidden links between news articles in large news corpora. There are two main works that utilize the connecting the dots approach as the basis.

Connecting the dots approach proposed by Shahaf et al. (Shahaf and Guestrin, 2010) tries to ensure a coherent news chain. A coherent news chain is characterized with smooth transitions between all articles through the whole chain besides strong pairwise connection between consecutive articles. The problem is formalized as a linear program to put constraints for ensuring strong pairwise association and smooth transitions all along the chain.

An alternative method for connecting the dots between news articles is suggested by Hossain et al. (Hossain et al., 2011), which is implemented within the scope of this study. The method is based on Swanson's complimentary but disjoint hypothesis (Swanson, 1991). Swanson characterizes an ideal chain as one that satisfies a distance threshold between consecutive article pairs while does not oversatisfy the threshold between non-consecutive news articles. The method constructs the chain out of a concept lattice. Concepts represent closed termsets. An article's successors are selected from the concept with the largest termset that contains this article. The next article in the chain is determined with respect to two criteria: clique size (k), and distance threshold. k neighbors are determined with respect to Soergel distance at each step and A* search algorithm is run to find out the chain with the given endpoint.

3 Method

3.1 Methodology on the Evaluation of News Chains

In literature, there is no well-established methodology to measure the goodness of a given news chain. Two works propose different, independent mechanisms to evaluate news chains but they are not experimentally validated or compared with each other as part of a methodology. Moreover, there is a lack of ground-truth datasets in this area.

In an effort to establish such a methodology, we were in search of some ground-truth thus we collected already produced news chains by different works. Referring to the example chain provided by (Shahaf and Guestrin, 2010), we constructed that chain by searching the given article titles in the New York Times Portal. We named this chain as **Shahaf et. al. news chain**. As a control condition, we constructed another chain out of this by putting three copies of the fourth document (41070964.xml) in its place. We call this news chain **Shahaf et. al. control 1**.

As another published news chain, we referred to Alderwood story (ah Kang et al., 2009) provided as part of the VAST 2006 challenge (Whiting et al., 2009). The dataset (composed of 1182 documents) provides a ground-truth chain of length 19 (**VAST 2006 Challenge news chain**).

In addition to these news chains, three random news chains of equal length are produced from the New York Times Annotated Corpus (Sandhaus, 2008).

A news chain evaluation methodology needs methods to calculate some "goodness metrics". One such method quantifies a news chain with respect to its coherence by using a linear program (Shahaf and Guestrin, 2010). The linear program uses $kTrans$ and $kTotal$ as constraints to compute the $minedge$ objective value. Thus, our first goodness metric is $minedge$.

3.1.1 Minedge - Linear Programming Approach

The proposed linear program calculates two kinds of scores for every word in the chain. The first one is the activation score (act), which is the frequency difference of a word in two consecutive documents. As for the second; the initiation score (init), the difference between the activation scores of a word in consecutive document pairs is calculated:

$$act = freq(i+1) - freq(i) \quad (1)$$

$$\text{init} = \text{act}(i+1) - \text{act}(i) = \text{freq}(i+2) - 2\text{freq}(i+1) - \text{freq}(i) \quad (2)$$

The linear program makes word selection with respect to these activation and initiation scores. In other words, the linear constraints are defined in terms of activation and initiation variables. The first constraint variable $kTotal$ constrains the sum of initiation scores in the whole chain. The other constraint variable, $kTrans$ limits the sum of activation scores on individual document transitions.

In order to calculate the $minedge$ objective value, the activation score of a word in a document pair is weighted by the influence of that word in

connecting those two documents and the weighted activation scores are summed. The objective is to maximize this sum.

The influence of a word in connecting the document pairs is calculated based on the document-word bipartite graph. In this graph, documents and words are nodes and normalized word frequencies are the edge weights. The influence of a word w in connecting the document d_1 to the document d_2 is calculated by the use of the path that connects d_1 to d_2 over w:

$$influence(d_i, d_{i+1}|w) = p(w|d_i).p(d_{i+1}|w) \quad (3)$$

In the original implementation of the $minedge$ metric value, (Shahaf and Guestrin, 2010) state that they use random walk on the document-word bipartite graph to calculate the influence of words. On the other hand, we apply the formula in equation 3 for simplicity.

One of the key issues in this linear program is tuning parameters $kTotal$ and $kTrans$ to maximize the $minedge$ value. In order to determine the best parameter values, we plotted $kTotal$, $kTrans$, and $minedge$ values in 3D (Figure 1). As can be seen from the plot, the dark red areas represent the maximized values of $minedge$, thus, we selected $kTotal$ and $kTrans$ values belonging to those areas.

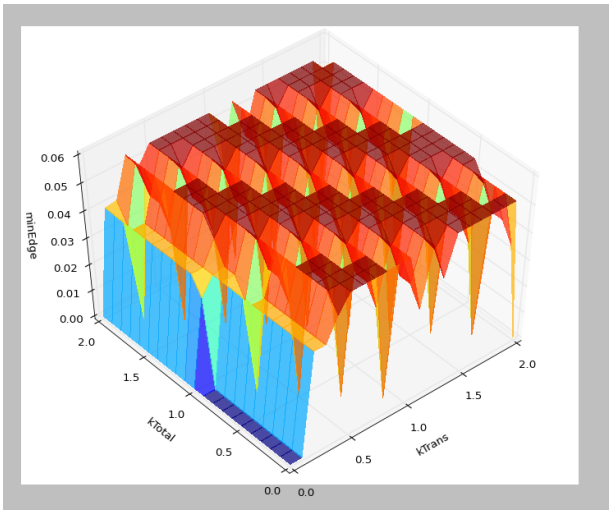

Figure 1: $minedge$ values with respect to $kTotal$ and $kTrans$.

We use the maximum $minedge$ value to quantify the coherence of a given news chain. Thus, in the given example the coherence score determined by the $minedge$ value is 0.053625.

3.1.2 Dispersion coefficient

Another proposed metric to measure the quality of news chains is $dispersion\ coefficient$ that is calculated based on Soergel distances (Hossain et al., 2012b).

Soergel distance between two documents is calculated considering all the words in the set of documents. The difference between the weights in the first and second document is summed for every word and this sum is normalized by the sum of the max. of those two weights for each word:

$$D(d_1, d_2) = \frac{\sum_t |w_{t,d_1} - w_{t,d_2}|}{\sum_t \max(w_{t,d_1}, w_{t,d_2})} \quad (4)$$

The weight of a word for a document is calculated using a variant of TF-IDF cosine normalization (Hossain et al., 2012a).

Soergel distance returns 0 for two identical documents and 1 for documents that do not overlap.

The dispersion coefficient metric-based quality evaluation is premised on Swanson's CBD (complimentary but disjoint) hypothesis (Swanson, 1991). The method computes a coherence score on the basis of Soergel distances between consecutive and non-consecutive pairs along the chain.

Dispersion coefficient is computed using the following formula:

$$V = 1 - \frac{1}{n-2} \sum_{i=0}^{n-3} \sum_{j=i+2}^{n-1} disp(d_i, d_j) \quad (5)$$

In the formula, $disp$ dispersion value inside the nested sums becomes positive when the angle between document pairs is above a specified threshold (or the distance between document pairs is below a specified threshold), otherwise it's 0. In the cases where it's positive, if the position difference of documents of pair is high, it affects by taking a higher value in other words it reduces the value of dispersion coefficient in a larger extent:

$$disp(d_i, d_j) = \begin{cases} \frac{1}{n+i-j}, & \text{if } D(d_i, d_j) > \Theta \\ 0, & \text{otherwise} \end{cases}$$

3.2 News Chain Construction

In this paper, we implemented the chain construction method suggested by Hossain et al.. The work constructs a concept lattice from the inverted index of documents and represents each closed termset

	Chain length	Minedge value
Shahaf et. al.	8	0.02
Shahaf et. al. control	10	0.018
VAST 2006 Challenge	19	0.0064
Random news chain 1	8	0.0055
Random news chain 2	8	0.0036
Random news chain 3	8	0.0023

Table 1: Minedge metric values.

	Chain length	Threshold 0.22	Threshold 0.25
Shahaf et. al.	8	0.836111	0.483333
Shahaf et. al. control 1	10	0.984375	0.615625
Shahaf et. al. control 2	10	0.829167	0.600893
VAST 2006 Challenge	19	0.862132	0.571804
VAST lattice all avg.	7.5	0.873	0.401
Random news chain 1	8	1.0	0.394444
Random news chain 2	8	0.883333	0.316666
Random news chain 3	8	1.0	0.711111

Table 2: Dispersion coefficient values.

by a unique concept. Each concept has terms as extents and documents as intents. We used CHARM-L algorithm (Zaki and Hsiao, 2005) to generate this concept lattice structure.

In order to generate promising candidate chains out of this lattice, an initial document has to be determined. The algorithm then proceeds by looking for the largest extent size concept that includes this initial document in its intent set. After that; inside this concept, candidate chains are sought using local neighborhood-based search. The search heuristic is defined by two criteria. Clique size determines the maximum number of neighbors to evaluate at each stage whereas distance threshold criterion makes a selection out of them.

As we worked with the VAST 2006 Challenge dataset, we selected the start document as the initial document of the VAST 2006 Challenge ground-truth chain. Then, we worked with 3 clique-size and 10 clique candidates (a total of 10 3 cliques) in order to find a good set of successors. As a result, we created all candidate chains from the VAST dataset.

3.3 Experimental Results

We calculated our goodness metrics for the news chains in our experimental design. In Table 2, $minedge$ metric values are shown. When we look at the obtained values, we observe that ground-truth chains have higher values than the randomly generated ones. Additionally, **Shahaf et. al. control 1** gets a lower value compared to its original chain. Thus, the results support the claim that $minedge$ metric value behaves in a correct and consistent way in measuring the coherence of given news chains.

As a second part of our experiment, we computed the dispersion coefficient values for all the chains by fixing the Soergel distance threshold value as 0.22 and 0.25 respectively.

The obtained dispersion coefficient values do not seem to work well in the quality assessment of news chains. First of all, **Shahaf et. al. control 1** has a higher score than **Shahaf et. al.**. **Shahaf et. al. control 1** has three copies of the fourth document and it is desired to have a lower dispersion value since repeating exactly the same news will not contribute to chain coherence. However, the dispersion coefficient approach does not consider the repetition of documents as a semantic parameter, it simply penalizes nonconsecutive documents which over-satisfies the distance threshold as a document pair. However, for a given chain we cannot know how close the nonconsecutive documents are beforehand. In order to verify this idea, we added a second control condition (**Shahaf et. al. control 2**) in which we repeated the fourth document at the beginning, middle, and end points, which resulted in lower dispersion coefficient values since non-consecutive identical document pairs have the distance value 0 and penalty is higher due to higher index differences. Moreover, random chains (Random news chain 3) can get comparable higher scores for this measure.

As for the lattice-based news construction algorithm, the average score for all the constructed chains does not make an important difference when compared with the VAST 2006 Challenge ground-truth chain.

4 Conclusion

The first goodness metric, $minedge$ gives correct and consistent results. However, the dispersion coefficient values fail to evaluate the "goodness" of given news chains. The reason can be attributed to disregarding the consecutive document pairs in the calculation of the coefficient value. Because penalizing with respect to far away documents in the chain is necessary but not sufficient condition for a chain definition. At the same time, strong pairwise association must be guaranteed.

When it comes to the lattice-based news chain construction algorithm, extensive experimental validation is needed. Moreover, alternative path traversal heuristics can be adapted to the constructed lattice to produce coherent news chains.

Acknowledgments

This paper is based on work supported by the Scientific and Technological Research Council of Turkey (TÜBİTAK) under contract number 114E784.

References

Youn ah Kang, Carsten Görg, and John T. Stasko. 2009. Evaluating visual analytics systems for investigative analysis: Deriving design principles from a case study. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology, IEEE VAST 2009, Atlantic City, New Jersey, USA, 11-16 October 2009, part of VisWeek 2009*. pages 139–146. https://doi.org/10.1109/VAST.2009.5333878.

M. S. Hossain, J. Gresock, Y. Edmonds, R. Helm, M. Potts, and N. Ramakrishnan. 2012a. Connecting the dots between PubMed abstracts. *PLoS ONE* 7(1):e29509.

M Shahriar Hossain, Christopher Andrews, Naren Ramakrishnan, and Chris North. 2011. Helping intelligence analysts make connections. *Scalable Integration of Analytics and Visualization* 11:17.

M. Shahriar Hossain, Patrick Butler, Arnold P. Boedihardjo, and Naren Ramakrishnan. 2012b. Storytelling in entity networks to support intelligence analysts. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, USA, KDD '12, pages 1375–1383. https://doi.org/10.1145/2339530.2339742.

E. Sandhaus. 2008. The New York Times Annotated Corpus. *Linguistic Data Consortium, Philadelphia* 6(12).

Dafna Shahaf and Carlos Guestrin. 2010. Connecting the dots between news articles. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, July 25-28, 2010*. pages 623–632. https://doi.org/10.1145/1835804.1835884.

Don R. Swanson. 1991. Complementary structures in disjoint science literatures. In *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, USA, SIGIR '91, pages 280–289. https://doi.org/10.1145/122860.122889.

M.A. Whiting, Chris North, Alex Endert, J. Scholtz, J. Haack, C. Varley, and J. Thomas. 2009. Vast contest dataset use in education. In *Visual Analytics Science and Technology, 2009. IEEE VAST 2009.*. pages 115 –122. https://doi.org/10.1109/VAST.2009.5333245.

Mohammed J. Zaki and Ching-Jui Hsiao. 2005. Efficient algorithms for mining closed itemsets and their lattice structure. *IEEE Trans. on Knowl. and Data Eng.* 17(4):462–478. https://doi.org/10.1109/TKDE.2005.60.