Summary

Connecting the Dots Between News Articles

Dafna Shahaf
Carnegie Mellon University
dshahaf@cs.cmu.edu

Carlos Guestrin
Carnegie Mellon University
guestrin@cs.cmu.edu

Presenter:
Monika Mitrevska

Opponent:
He Niu

Supervisors:
Maya Ramanath
Ralf Schenkel

Motivation

Nowadays the users are exposed to huge amount of information published on internet every day. Keeping up with so much information can be confusing and misleading. The users can easily miss the big picture. The general motivation for this paper is the information overload and the necessity to bring closer the information users need

Problem

One of the biggest problems in today's society is the problem of extracting useful information from the available large datasets. This problem also applies to the news domain, especially for the complex stories covering more than one event. An example for the possible difficulties with this kind of stories is when we have two events, but we are not really sure how are they connected. The hidden connections (events that happened in between) are what the user usually wants to discover. One example mentioned in the paper was the question how the reforms in the health care system are connected with the credit crises. With the standard information retrieval, using the available search engines, finding useful information for this type of stories can be complicated and time consuming for the users. This happens because these complex stories are really hard to be summarized by a single query strings. The alternative solution that the authors of the paper propose automatically connects the dots between the events, providing a structured and easy way to navigate within the topics and to discover hidden connections between them. The system focuses on the news domain. The general idea is for the user to pick and present to the system two news articles, then the system returns a chain of articles that connect the events from the given ones, forming a coherent story between them.

Contribution

The problem of connecting the dots is novel. By formalizing the characteristics of a good story through the notion of coherence, by formalizing the influence without link structure and by providing an efficient algorithm for connecting two fixed end points while maximizing the chain coherence the paper opens a new perspective in the information extraction and presentation domain.

General overview

The main idea of this paper is to connect the dots between two news articles. The user picks two news articles that describe the events he wants to connect. The goal of the methods presented in the paper is to find the best path between these two articles. At the end the user gets a set of articles that are covering the story between the first one and the last one. After reading this set of articles, the user should be able to understand the hidden connection between the articles and understand the general story. In order to be able to do that, the user must be provided with articles that form together a well structured and coherent story.

Formalizing story coherence: The formula:

$$Coherence(d_1,...,d_n) = \min_{i=1...n-1} \sum_w Influence(d_i, d_{i+1} \mid w) 1(w \ active \ in \ d_i, d_{i+1})$$

is defining the coherence for a given chain of articles.

Every chain is as strong as its weakest link. This formula is taking the value, the strength, of the weakest link as a measurement for the coherence. The strength of the transition (link) between two documents is measured as a sum of the influence of every active word in that transition (link). The intuition behind it is that they measure the story flow between two documents by summing the influence of the words. If more words are influential for the transition that means that bigger part of the story continues in the next document.
Measuring the importance and not the appearance of the words in the documents allows counting words that are not physically present in the documents, but still are important for the transition. It also allows giving different important levels to the words. Introducing the activation level for every word and transition prevents from jitteriness – topics that appear and disappear along the chain.
The problem of maximizing the value of the coherence is formalized and solved as a Linear Programming problem. In the LP formulation are given n chronologically ordered documents $d_1 \ldots d_n$. For every document i and word w the variable $word\text{-}active_{w,i}$ is measuring the activation level of word w during the transition from d_i and d_{i+1}. The variable $word\text{-}init_{w,i}$ indicates the initialization level of w in d_i. The LP solution the paper presents, has three main parts:
Smoothness: This module makes sure that every word is initialized at most once and that if one word is active in a transition that means it was active in the previous one as well, or it was just initialized.

$$\forall w \sum_i word - init_{w,i} \le 1$$

$$\forall w,i \ \ word - active_{w,i} \le word - active_{w,i-1} + word - init_{w,i}$$

$$\forall w \ \ word - active_{w,0} = 0$$

Activation Restrictions: This module limits the total number of active words and the number of words that can be active during a single transition. These two restrictions are used to control the length of the activation segments.

$$\sum_{w,i} word - init_{w,i} \le kTotal$$

$$\forall i \sum_w word - active_{w,i} \le kTrans$$

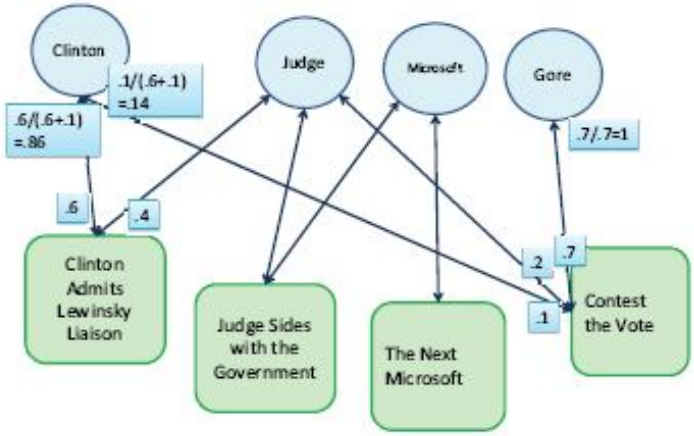
Objective: The influence is calculated for every link and the minimum (the weakest link) is saved into the variable *minedge*. The objective is to maximize this variable

$$\forall i \;\; \text{minedge} \le \sum_{w} \text{word} - \text{active}_{w,i} * \text{influence}(d_i, d_{i+1} \mid w)$$

$$\forall i, w \;\; word - active_{w,i}, word - init_{w,i} \in [0,1]$$

The definition of the objective and the restrictions in more details can be found in the paper.

Measuring influence without links: *Influence(d_i,d_j|w)* measures the influence of the word w in the transition from the document d_i to the document d_j. This measurement is used to find the strength of a possible link between two articles. Because in the model considered in the paper there are no edges between the documents, the authors explore a different notion of influence, taking the words into account. They first take all the preselected documents, then extract all the named entities and noun phases using CopernicSummarizer[1] tool. After that, a bipartite directed graph G=(V,E) is created, where V_D U V_W corresponds to documents and words. The importance of each word given by the CopernicSummarizer is used as an edge weight. As it is shown on the image below, the document–to-word and word–to-document weights are normalized over the words and the documents respectively. To calculate the influence of the word w in the transition between the documents d_i and d_j with this graph, they first compute the stationary distribution for random walks from d_i to d_j. To calculate the influence of w, they make w to be a sink node in the graph and again calculate the stationary distribution for the random walk from di to dj. The difference between these two distributions represents the influence of the word w. Intuitively, measuring the probability to get from di to dj with and without w tells actually how important is w for this connection.



Finding a good chain. The problem of finding the coherence of a chain is generalized into a problem of finding a good chain. Solving this problem has the same approach as finding the coherence, with the difference that here the nodes and the links are unknown, so they are all taken as possible candidates.

This problem, same as the previous one, is formulated and solved as a linear program. The objective of the linear program is to maximize the coherence for all possible chains under three sets of restrictions: Chain restrictions, Smoothness and Activation restriction. The Chain Restriction is a new module, compared with the LP for

[1] Copernic, http://www.copernic.com

finding the coherence of a chain and it ensures a proper chain. It ensures that the chain will start and end with the given articles, that the chain will have K nodes ordered chronologically and K-1 edges.

$$node-active_1 = 1, node-active_n = 1$$

$$\sum_i node-active_i = K,\ \sum_i next-node_{i,j} = K-1$$

$$\sum_i next-node_{i,j} = node-active_j \quad j \neq s$$

$$\sum_j next-node_{i,j} = node-active_i \quad i \neq t$$

$$\forall_{i \geq j}\ next-node_{i,j} = 0$$

$$\forall_{i<j<k}\ next-node_{i,j} \leq 1 - node-active_k$$

The complete formulation of the linear program can be found in the paper.
The result of the linear problem is a fractional directed flow starting with the first given, ending with the second given article In order to find the best path between them, the paper present a randomized rounding schema with proven guarantees.

Scaling up: The defined LP problem has $O(|D|^2*|W|)$ variables, so it is impossible to be solved for large number of articles. The paper suggests that the bipartite graph mentioned above can be used to choose the best suitable subset of articles. They do that by running random walks from the beginning and the ending article and take the highest ranked ones. Because the random walks start from both articles, they believe that the articles that are recently reached from both sides are most important and going to be highly ranked.

The process of calculating the influence is also speeded up by using one set of random walks for all w. They simulate the random walks in the original graph for every document, and remember the words encountered, then, when taking every word as a sink, they use the same random walks, just without that word.

Interaction models. Having the structured nature of the chain, the authors argue that makes sense to go beyond the standard methods for evaluation in information retrieval and explore more expressive forms of interaction. Instead of the standard way of letting the users to revise their queries they present in the paper two different types of user feedback: refinement of a chain and tailoring to the user interest. The refinement of the chin allows the user to pick an article they don't like in the chain, or to pick a place in the chain where they want more information (one additional article). The system tries all the possible replacements/insertions for the picked article/place and returns the best one.

The feedback that incorporates user's interests allows the users to increase the importance of some words and with that, as a result they would get a refined chain in the direction they choose. In order to take the user's feedback into account, they introduce one more variable in the system – importance weight π_w.

Evaluation

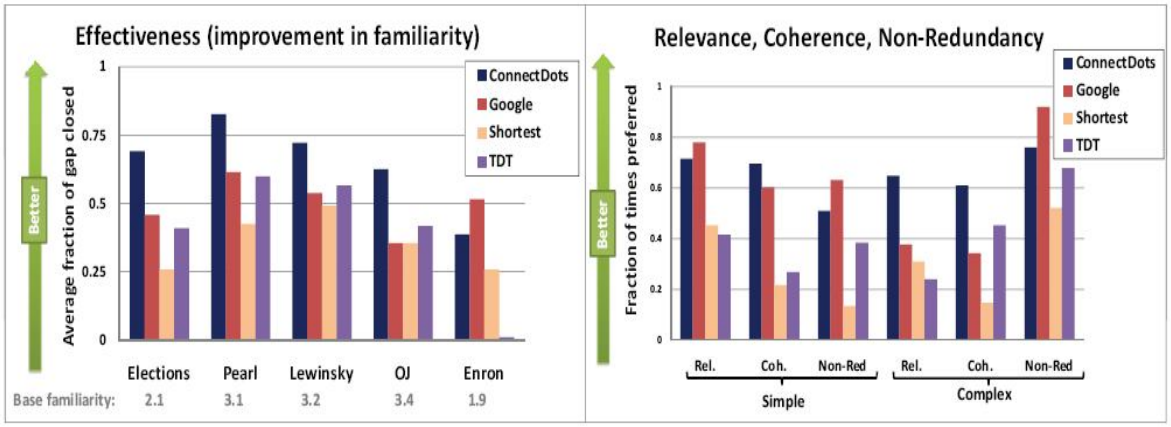
Since there is not available labeled dataset suitable for the task the paper is solving, the authors decide to run their methods on real data, on news articles from New York Times and Reuter's dataset (1995 – 2003)

They pick four topics and use four methods to produce the chains. The topics are: OJ Simpson trial; the impeachment of Clinton; the Enron scandal and September 11th and the methods: Connecting the dots, Shortest-path, Google News Timeline and Event threading. In the evaluation participated eighteen users. At the beginning, the source and target articles were presented to the participants, and they were asked to indicate their familiarity on these articles on scale of 1 to 5. After reading the chains of articles they were asked to indicate the relevance, coherence and redundancy for every chain, and they were also asked to measure the familiarity about the stories, now, after reading the chains. They measured the effectiveness of the chains by the fraction of the familiarity gap closed. The results are shown on the image below. The graph on the left shows the effectiveness of each of the methods for every story. Except for the Enron story, for every other, connecting the dots method was the most effective one. The authors argue that for simple stories, for which there are not so many articles available, picking K equally spaced documents was sufficient for most of the participants. The Enron's story is of this kind.

The right side the graph represents the results for the relevance, coherence and non-redundancy.

The results are here divided into results for simple stories (focus around same event) and result for complex stories (connected through one or more events). In general connecting the dots method has better results on the complex stories compared to the results on the simple ones. For both cases connecting the dots method is the best at creating a coherent story. This means that they have succeeded in maximizing the coherence of the story.

When it comes to the problem in having redundant articles they argue that the relevance and the redundancy are proportional and that there has to be a trade – off between them.



Discussion

In my opinion this paper is very well written and the problem and the solution are very well elaborated. Even in the discussion session after the presentation there were only few discussions questioning the paper.

One of the discussion points was about the application of this system having in mind the processing time for every chain. In the evaluation part they mentioned 10 minutes processing time per chain, which in my opinion, is too long for “every day” information retrieval. Even more, the beginning and the ending point of the chain depend on the user and the quality of the resulting chain depends on this two articles. If the user wants to refine his choice several times, than using this system can get more complicated and time consuming than the standard browsing.

Another area where there is maybe a possible place for improvement is the redundancy. This paper justifies the bad non-redundancy performance with the relevance. They never mention if they even consider the redundancy when creating the chain. In the discussion was mentioned that there is a way to improve the non-redundancy with a redundancy check. In this case we cannot be sure if the problem can still be solved as LP problem, and solving an information retrieval problem as an LP problem makes this paper novel and relevant.

One can also argue if it is enough to take into consideration only the plain set of words or should take the semantics of the sentences into account as well. Taking the semantics into account can open place for improvement of the system. The system then, for example, would be able to distinguish between positive and negative connotation of the events. Also maybe would be able to protect the chain from spamming articles.

Future work: As their future work, the author mention that they want to see how the system performs with events that are less popular and less covered. Also they say they want to introduce richer forms of input and output for more complex task. One example would be a Roadmap output: set of chains that will cover more aspects of the story.