Assignment 4: The Last Bits
Storylines

Shrestha Ghosh (2567717, s8shghos@stud.uni-saarland.de)

September 20, 2018

1 Introduction

Event tracking has become quite cumbersome given the huge amount of data being continuously generated online. It is very easy to get trapped in details and lose the big picture. With the multitude of interconnecting link between content, forming a coherent understanding of the development of an event becomes very difficult. In this report we look into two research work which tackle this problem. The end goal is to automatically provide to the user a structured and easy navigation within a topic with the bonus of discovering links which are not directly established. We shall closely analyze how the system is modeled, whether the experiments are fair and the evaluations sufficient to support the results.

2 Connecting the Dots Between News Articles

Shahaf and Guestrin [1] present a automatic event tracking system which returns a chronologically ordered chain of articles linking two news events where the returned articles aim to provide a consistent coverage of topics/events from the starting event (s) till the end (t). Given a set of documents, the problem boils down to finding the chain of news articles linking s and t and defining a score function in order to select the best chain.

Since there is no predefined structure like graphs to capture the links between the news articles (just a collection of articles), the authors define a correlation measure to compute the "influence" of a document on another document. A bipartite graph is formed, where one set contain the document nodes and the other set comprises word nodes. Each word node is connected to the document nodes in which it occurs with a bi-directional edge. The document-to-word edges weighed with tf-idf (or similar word frequency measure) of the words in the document. Conversely, the word-to-document edges are weighted with the tf-idf score of the word node normalized over documents it occurs. This setup is used to mimic random walk in two scenarios. The influence of a document (d_i) on another document (d_j) w.r.t a word w is the difference in the stationary distribution of d_j on the full graph and on the modified graph where the node w has no outgoing edges for a walk starting at d_i. This w node is an absorbing node, therefore, any document d_j highly influenced by w will show a decrease in its stationary probability since it is no longer reachable from w. This ad-hoc influence measures the decrease in the average number of visits to d_j from d_i when w becomes a sink node. However, there is no sense of direction of time. This means that a the same event covered by different journalists at one time will have higher influence measure than an article which covers the "developments" in the event at a later point in time.

There are restrictions on active words based on the assumption that there are few event related words which recur throughout the chain. Firstly, at most a fixed number of words can be activated in the entire chain of 6 to 7 articles. Secondly, each word activation is restricted to occur only once. This means that there are no breaks in the occurrence of a words across the chain once it starts and it may not appear again once it is absent in an article. The third restriction is on the number of words active between adjacent articles which is set to 4. These restrictions may make the system susceptible to noisy and redundant data. The problem of redundancy arises as pointed out previously due to the absence of a notion of direction of time such that parallel news with highly overlapping influential words are preferred over developmental news.

Chains are susceptible to noise for lesser known events. In their paper, the authors some very famous events with highly specific topic. In real scenarios, however, people might search for more localized (geographically or professionally) or unfocused news. For example if a user wants to track "Fifa World Cup 2018" to "Mesul Ozil retires from national team", the chain could digress to many aspects of the World Cup at the beginning.

There are fairness issues with regard to the evaluation method chosen by the authors to demonstrate the results of their in comparison to other systems. The Connecting-Dots system runs an optimization which restricts the size of the chain. Thus, for the specified chain length, the system finds the most coherent chain. For other systems there is no such restriction (shortest path, Google news and Event Threading). Therefore, selecting equally spaced articles may not do justice in maintaining the coherence that the original chain has. The authors run their system and shortest-path technique on a reduced dataset with the "hope" that highly ranked documents in a random walk in an initial bipartite graph containing all data are also the the ones most frequented. Articles are iteratively and selectively added if the optimization returns a weak chain, however, the process of identifying a weak chain and subsequent article selection is not very well explored.

A more comprehensive evaluation should include the performance of Connecting-dots on a dataset built from multiple news sources (resonating the actual motivation of information explosion). The original idea is to prevent users from getting lost in information. If the performance is tested on highly refined and few sources (Reuters and New York Times) it is not comparable to say the Google News which sifts through multiple sources covering time-parallel data and then generating a news timeline. The user interest study is presented in one way and conducted in another. The user interest test was aimed to test how the system incorporates user interests to provide a more tailored result. The actual experiment, however, required the user to identify perceivable "user interests" shifts in the new chain. This may be because the authors wanted to restrict the number of user interests to the activated words only. A better way would have been to present the activation words (since there aren't many) to the user and collect their top k-preferences. Then present the resulting chain and get the user satisfaction rating. This would be more in tune with the goal of tailored results.

3 Metro Maps of Science

Shahaf et al. [2] develop upon the idea of coherent chains between news articles to adapt top the domain of scientific publications. While the objective of providing a structured and easy navigation remains the same, the motivation is different. Previously we were more interested in the timeline of a topic from a start till an end event [1]. In the current domain of scientific studies, the motivation is to identify the different lines of research stemming from a particular topic of interest. Hence, given a topic we would like to know about the developments in the field with respect to different lines of research stemming from the topic. Specifically the authors aim to provide a "metro map" tracing the research lines concerning a specific field/topic. The authors formalize the notions of coherence, coverage and connectivity specific to the domain of scientific publications.

The order of the constraint classes - first coherence, second coverage and finally connectivity - is very definitive aspect of the system. Given a set of documents the space of possible metro maps is exponentially large. The coherence constraint which aims to identify all coherent chains in the document space heuristically reduces the search space. Since the rest of the solution is built upon the chains resulting from coherence constraints it becomes essential that the heuristics and approximations used are reliable. Once we have coherent chains tracing different lines of research within the document space, the coverage constraint is applied to retain only those chains which maximize the coverage of the search topic. This means that we are looking for chains which include diverse research lines rather than research lines concentrated in a specific area. The final constraint of connectivity rewards those maps which have higher number of interconnected chains, *i.e.,* primarily divergent lines of research which influence each other. We should note that the metrics of coherence, coverage and connectivity cannot be all optimal at once. Highly connected chains are not necessarily well connected similarly, highly coherent chains may be sparsely interconnected. In the same fashion, there is a trade-off between coverage and connectivity. The order of the constraints signifies the degree of trade-off. Hence, the metro map should primarily contain coherent paths, which cover diverse topics and there should be some connectivity between different paths (the authors define a soft notion of intersecting paths).

The systems considers data on a predefined area of interest to generate metro maps of documents published in a particular conference or journal, and therefore the system does not have to worry about noisy data (like unplublished archive papers). The redundancy is taken care of by the coverage constraint. Since the usage of maps is very user specific, evaluation metrics like precision and recall have to be taken with a grain of salt. Also, the evaluation in Wikipedia is unfair and doomed from the start. Metro maps and Google Scholar contain clean and high quality data, Metro Maps more so because it works on documents of the specified domain published in the top conference. A more thorough evaluation should have also accounted the documents from the conferences/journals mentioned in the Wikipedia article for the topic and presented the same to the user - in that case the system maintains at least some parity with Google Scholar database. An interesting result would be to evaluate the performance of Metro Maps on the documents returned by Google Scholar for a topic (as researchers this seems more practical, because we do not want to limit to papers in one particular journal). Similarly, we could gauge how Metro Maps performs on the references returned from Wikipedia article on a topic. To make it more robust we could include all references from Wikipedia pages which are k External links away from the original topic. In this way we would have three outputs of the Metro Maps on three different domains and perform evaluation metrics without having to be heavily dependent on recruiting participants with similar background knowledge and an expert judge.

4 Conclusion

While the papers incorporate random walk properties in their system, the absence of edges and actual computation of link prediction calls for a very ad-hoc solution to the problem at hand. The authors completely dismiss the co-citation links and coherent chains are dependent on the notion of influence based on topics. Since citation is an important part in tracing lines of research it must have some weight in determining coherent chains. Or, does Metro Maps' $Influence$ measure inherently account for this? We have seen that the idea of connectivity takes into account interactions spanning "affirmation, criticism, contrast, methodology, and related work". Evaluating the extent to which maps overlap with a citation link graph (edges from a document citing previous work) or approaching the problem by reducing a citation link graph to a metro map might also lead to interesting results.

References

[1] Dafna Shahaf and Carlos Guestrin. Connecting the dots between news articles. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 623–632. ACM, 2010.

[2] Dafna Shahaf, Carlos Guestrin, and Eric Horvitz. Metro maps of science. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1122–1130. ACM, 2012.