Abstract

Modeling influence of entities in networked data is an important problem in information retrieval and data mining. Popular algorithms such as PageRank capture this notion of authority by analyzing the hyperlink structure, but they ignore the topical content of the document. However, often times, authority is topic dependent, e.g., a web page of high authority in politics may be an unknown entity in sports.

In this work, we describe a new model called TopicFlow, that combines ideas from network flow and topic modeling, and captures the notion of topic specific influences of hyperlinked documents in a completely unsupervised fashion.

We show that on the task of citation recommendation, the TopicFlow model, when combined with TF-IDF based cosine similarity, outperforms several competitive baselines by as much as 6.4%. We also present some qualitative visualizations to demonstrate the expressive power of the new model.

1 Introduction

Finding authoritative entities in hyperlinked data such as the world-wide-web, academic literature, blogs, social media, etc. is an important problem in information retrieval and data mining. Popular algorithms such as PageRank [7] and HITS [5] have been very effective in identifying influential documents by analyzing the hyperlink structure of the corpus. Although quite successful, one of the shortcomings of such link based algorithms is that they completely ignore the topics discussed by these documents. In reality, the influence\(^1\) of an entity is highly dependent on the topical context. For example, an entity that is highly influential on the topic of politics may be unknown in sports.

Although this problem was addressed by Topic Sensitive PageRank (TSP)[4], it requires that the topics are prespecified and that a few labeled documents for each topic are available in the corpus. However in most corpora, documents are unlabeled and topics of the corpus are unknown.

In the last decade, there has been tremendous progress in automatically learning the topics of an unlabeled corpus, and soft labeling the documents with these topics, in a completely unsupervised fashion [1]. This approach, called the topic modeling approach, is a natural candidate in terms of our objective of unearthing topics and modeling topical authority simultaneously. More recently, many researchers have focused their efforts on extending topic models to include hyperlink structure as part of the model [9, 6, 2, 3]. Although these models were successful in modeling topical correlations across relational links, none of them captures the notion of global topic-specific influence of documents.

2 TopicFlow model

We consider a directed graph \(G = (V, E)\) called the influence graph, where \(V = \{d_i\}_{i=1}^M\) is the set of all \(M\) documents in the corpus. We define \(K\) directed edges \(e_{ij}^{(k)}\) from document \(d_i\) to \(d_j\) if \(d_j\) cites \(d_i\) where \(K\) represents the number of topics in the corpus. Note that this is opposite to the direction of citation and represents the direction of flow of influence. Each directed edge \(e_{ij}^{(k)}\) acts

\(^1\)In the rest of the paper, we will use the words ‘influence’ and ‘authoritativeness’ interchangeably.
as an infinite capacity channel for “flow of influence” from document \(d_i\) to document \(d_j\) on topic \(k\). The actual flow for topic \(k\), represented by \(f^{(k)}_{i,j}\), is restricted to be non-negative and in the same direction as the edge, and represents the influence of document \(d_i\) on document \(d_j\) on topic \(k\).

We now define an augmented graph \(G' = (V', E')\) such that \(V' = V \cup \{s, t\}\) where \(s\) and \(t\) represent a fictitious source and sink respectively. Similarly, we define the augmented set of edges \(E'\) as follows:

\[
E' = E \cup \{e^{(k)}_{s,d} \mid d \in V; k \in K\} \cup \{e^{(k)}_{d,t} \mid d \in V; k \in K\}.
\]

We assume that a unit flow arises out of the source \(s\), flows into the network through the augmented edges from \(s\) and reaches into the sink \(t\) through its augmented edges, as shown in Fig. 1.

Conceptually, the source represents the flow of influence of topics and ideas that are unexplained by the set of hyperlinks in the graph \(G\). Another reason for defining the source is to model the influence of all ideas from the past with respect to the documents in \(G\). Similarly, the sink can be thought of as an abstraction to the flow of ideas from the documents in \(G\) to the future.

We assume that the flows are balanced for each document-topic pair \((d, k)\) as follows:

\[
\sum_{i \in \text{Pa}(d)} f^{(k)}_{i,d} = \sum_{j \in \text{Ch}(d)} f^{(k)}_{d,j}
\]

(2)

where \(\text{Pa}(d)\), read as parents of \(d\), represents the set of vertices that have an outgoing edge into \(d\) in the augmented graph \(G'\) and \(\text{Ch}(d)\), read as children of \(d\), is the set of vertices that have an incoming edge from \(d\) in \(G'\). At each document, we allow the incoming flow on each topic to be split arbitrarily across its children with the exception that a uniform fraction of the flow always flows into the sink as shown below:

\[
f^{(k)}_{d,t} = f^{(k)}_{i,d} / (|\text{Ch}(d)|)
\]

(3)

where \(f^{(k)}_{i,d}\) is short for net incoming flow on topic \(k\) into \(d\) and is equal to \(\sum_{i \in \text{Pa}(d)} f^{(k)}_{i,d}\)

and \(|\text{Ch}(d)|\) is the size of children of \(d\) in \(G'\). This condition naturally satisfies the flow balance condition for documents that have no outgoing edges to other documents. For these documents, \(|\text{Ch}(d)| = 1\) (since the sink is their only child), hence the entire incoming flow into such documents on a topic flows entirely into the sink \(t\).

At the source, we have the following source flow balance constraints: \(\sum_{d=1}^{M} f^{(k)}_{s,d} = 1\).

For generating the words in the document, we use a process identical to LDA [1], with one significant difference. Instead of using a Dirichlet prior over the multinomial topic distribution over topics \(\theta_d\), we assume that the distribution is proportional to the topical inflows into the document:

\[
\theta^{(k)}_d = (f^{(k)}_{i,d}) / (\sum_{k'=1}^{K} f^{(k')}_{i,d})
\]

(4)

This is a key assumption that ties the network flow model with the topic model that generates words on specific topics in each document. The rest of the generative process for generating the document is identical to LDA [1], and is not reproduced here owing to space constraints. Accordingly, the

\[\text{By definition, } \forall d \in V: s \in \text{Pa}(d) \text{ and } t \in \text{Ch}(d).
\]

\[\text{We will also use the notation } f^{(k')}_{d,t} \text{ to represent } \sum_{j \in \text{Ch}(d)} f^{(k')}_{d,j} \text{, the net outgoing flow from a document } d.
\]

\[\text{Please see steps 3(a) and 3(b) on page 4 in the original LDA paper [1].}\]
observed data log-likelihood for the whole corpus is given by:

\[
\log P(\mathbf{w} | \beta, \mathbf{f}) = \sum_{d=1}^{M} \sum_{n=1}^{N_d} \log(\sum_{k=1}^{K} \beta_{k w_n} \theta^{(k)}_d),
\]

where \(\beta_k\) is a multinomial over the vocabulary for topic \(k\) and \(N_d\) is the length of document \(d\).

In this framework, a dynamic interplay between citations and words determines the influence of a document on a given topic. If a document \(d\) discusses a topic \(k\) in great detail, it must have a high \(\theta^{(k)}_d\) to explain its words, which induces higher topical flow \(f^{(k)}_{d, t}\) into the document relative to other topics. This will in turn flow out into documents citing \(d\), thus spreading its influence. On the other hand, if the document \(d\) is highly cited by other documents on topic \(k\), it must have heavy outgoing flow \(f^{(k)}_{d, s}\) on that topic, which will in turn induce high \(\theta^{(k)}_d\) for that topic, resulting in the assignment of many words in \(d\) to that topic.

We claim that the net outgoing flow from document \(d\) on topic \(k\) into its children not including the sink \(t\), denoted by \(f^{(k)}_{d, t} = \sum_{j \in \text{Ch}(d) - \{t\}} f^{(k)}_{d, j}\), captures the global influence of document \(d\) on topic \(k\). The topical flow balance conditions at every vertex ensure that the incoming flow at every document depends on the ‘supply’ of flow from vertices ‘upstream’ of \(d\). Likewise, the outgoing topical flow at each document is influenced by the ‘demand’ for topical flow by vertices in its ‘downstream’. Hence the flow parameters learned by the model capture truly global influence.

### 3 Learning and Inference

The optimization problem for our model is the following:

\[
\begin{align*}
\max_{\mathbf{f}, \beta} \mathcal{F}(\mathbf{f}, \beta) &= \log P(\mathbf{w} | \beta, \mathbf{f}) - \frac{1}{2} \lambda \left( \sum_k ||\mathbf{f}^{(k)}_s||^2 + \sum_d \sum_k ||\mathbf{f}^{(k)}_d||^2 \right) \\
\text{s.t.} \quad \forall d \in V' - \{s\} \sum_{i \in \text{Pa}(d)} \mathbf{f}^{(k)}_{i,d} = \sum_{j \in \text{Ch}(d)} \mathbf{f}^{(k)}_{d,j} \text{ and } \sum_{d=1}^{M} \mathbf{f}^{(k)}_{s,d} = 1
\end{align*}
\]

where \(\lambda\) is the coefficient of regularization and \(\mathbf{f}^{(k)}_s\) is a vector consisting of all the flows from the source on topic \(k\), while \(\mathbf{f}^{(k)}_d\) is the vector of all flows from document \(d\) to its children on topic \(k\). L2 regularization is introduced to ensure that all the flows remain small and as close to uniform as possible unless required by the data, and also to promote identifiability of the solution. We eliminate the equality constraints in Eq. 6 using the following equivalent flow balance condition:

\[
\forall j \in \text{Ch}(d) \mathbf{f}^{(k)}_{d,j} = \mathbf{f}^{(k)}_{d,\text{d,j}} \text{ s.t. } \sum_{j \in \text{Ch}(d)} \mathbf{f}^{(k)}_{d,j} = 1 \text{ and } \forall j \mathbf{f}^{(k)}_{d,j} \geq 0
\]

where the new multinomial variable \(\psi^{(k)}_{d,j}\) for each document-topic pair \((d, k)\) determines how the net incoming flow into \(d\) on topic \(k\) is split among its children \(\text{Ch}(d)\).

At the source, we define \(\psi\) as follows:

\[
\mathbf{f}^{(k)}_{s,d} = 1 \cdot \psi^{(k)}_{s,d} \text{ s.t. } \forall k \psi^{(k)}_{s,d} \geq 0 ; \sum_{d=1}^{M} \psi^{(k)}_{s,d} = 1
\]

Using a variational posterior multinomial distribution \(\phi_{dnk}\) over topics for each position \(n\) in document \(d\), we can define a lower bound on the log-likelihood of observed data as follows:

\[
\log P(\mathbf{w} | \beta, \mathbf{f}) \geq \sum_{d=1}^{M} \sum_{n=1}^{N_d} \sum_{k=1}^{K} \phi_{dnk} (\log \beta_{k w_n} + \log \theta^{(k)}_d - \log \phi_{dnk})
\]

Maximizing the lower bound w.r.t. \(\phi_{dnk}\) and \(\beta_{k w_n}\) yields the following update rules:

\[
\beta_{k w_n} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dnk} ; \phi_{dnk} \propto \beta_{k w_n} \theta^{(k)}_d
\]


However, estimating the parameter $\theta_d$ is non-trivial since it involves the flow parameters as given by Eq. 4. Substituting this equation into Eq. 9 yields the following:

$$\log P(w | \beta, f) \geq \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dnk}(\log \beta_{kw} + \log f_{d,k}^{(k)} - \log(\sum_{k'} f_{d,k'}^{(k')}) - \log \phi_{dnk})$$

(11)

The only parameters that remain to be estimated are the flow parameters $f_{ij}^{(k)}$ for each edge $e_{ij}$ and topic $k$. However, instead of estimating the flow parameters directly, we estimate $\psi$’s, the flow splitting proportions using the relation in Eq. 7. Notice that the multinomial parameter vector $\psi_{d,k}$ enters into the lower bound in Eq. 11 only through the log likelihood terms for the children of $d$. Hence we only consider the observed data log-likelihood for $Ch(d)$ to optimize $\psi_{d,k}$:

$$\log P(w | \beta, f)_{v \in Ch(d)} \geq \sum_{v=1}^{N_v} \sum_{k=1}^{K} \phi_{vnk}(\log \beta_{kw} + \log f_{d,v}^{(k)} - \log(\sum_{k'} f_{d,v}^{(k')}) - \log \phi_{vnk})$$

(12)

We can now express $f_{d,v}^{(k)}$ in terms of $\psi_{d,v}^{(k)}$ as follows:

$$f_{d,v}^{(k)} = f_{d,v}^{(k)} + f_{d,v}^{(k)} = f_{d,v}^{(k)} + f_{d,v}^{(k)} \psi_{d,v}^{(k)}$$

(13)

where $f_{d,v}^{(k)} = \sum_{u \in Pa(d) - d} f_{u,v}^{(k)}$ is the total flow into document $j$ on topic $k$ from all its parents other than $d$. The second step in the above equation arises from the relation in Eq. 7. Although we have succeeded in eliminating equality constraints from the original problem in Eq. 6, we still have additional equality and inequality constraints expressed in Eq. 7 to guarantee that $\psi_{d,k}$ remains a multinomial vector. We handle these constraints by further using a multinomial logistic transformation as shown below:

$$\psi_{d,v}^{(k)} = \left\{ \begin{array}{ll}
(1 - \frac{1}{|Ch(d)|}) \sum_{v' \in Ch(d) - v} exp(\eta_{d,v}') & \text{if } v \neq t \\
\frac{1}{|Ch(d)|} & \text{if } v = t
\end{array} \right.$$ 

(14)

where the variable $\eta_{d,v}^{(k)}$ is now unconstrained. Note that the probability of outflow into sink is obtained from Eq. 3. At the source, we define the logistic transformation for $\psi_{s,d}^{(k)}$ in line with its definitions in Eq. 8 as follows respectively:

$$\psi_{s,d}^{(k)} = \frac{\exp(\eta_{s,d}^{(k)})}{\sum_{d' \in \Theta(u)} \exp(\eta_{s,d'}^{(k)})}$$

(15)

We can now substitute Eqs. 13 and 14 into Eq. 12 to optimize $\eta$’s of the documents directly in an unconstrained way. The final equations for derivatives of the objective function w.r.t. $\eta_{u,d}^{(k)}$ below, where $u$ is a parent document of $d$, i.e., $u \in Pa(d) - s$ are given by:

$$\frac{\partial}{\partial \eta_{u,d}^{(k)}} F(f, \beta) = \left( \frac{\phi_{d,k} - N_d}{f_{d,k}} - \sum_{d' \in Ch(u)} \psi_{u,d'}^{(k)} \left( \frac{\phi_{d',k} - N_{d'}}{f_{d',k}} - \sum_{d'' \in Ch(u)} \psi_{u,d''}^{(k)} \left( \frac{\phi_{d''} - N_{d''}}{f_{d'',k}} \right) \right) \right) \psi_{u,d}^{(k)} f_{u,d} f_{s,d} (1 - \frac{1}{|Ch(u)|})$$

(16)

where $\phi_{d,k} = \sum_{n=1}^{N_d} \phi_{dnk}$ and the second term above is computed only once for all $d' \in Ch(u)$.

Similarly, at the source, using the outflow relations in Eq. 8, and the logistic transformations in Eq. 15, we get the following equations for the derivative:

$$\frac{\partial}{\partial \eta_{s,d}^{(k)}} F(f, \beta) = \frac{\partial}{\partial \eta_{s,d}^{(k)}} \left( \sum_{d'} \sum_{k'} \phi_{d'nk'} (\log \beta_{kw} + \log f_{d'nk'}^{(k')} - \log(\sum_{k''} f_{d'nk''}^{(k'')}) - \log \phi_{d'nk'}) \right)$$

(17)

We optimize $\eta_{u,d}^{(k)}$ and $\eta_{s,d}^{(k)}$ by performing gradient ascent using the derivatives in Eqs. 16 and 17. At inference time, we keep the topic-word distributions, $\beta$, fixed at the learned values and learn the flow parameters for the test documents as done in training.
4 Experiments

The dataset we considered is the ACL anthology [8] dataset comprising full text and abstracts of all papers published in the annual Association of Computational Linguistics conference, over a period of over 30 years. For our experiments, we used the full text of 9,824 papers published before or in 2005 as the training set. There are 33,604 hyperlinks in total in the training set and some of the documents contain no incoming or outgoing hyperlinks. We used 1,041 abstracts of papers published in 2006 with no hyperlink information within this data, as the test set. However, we do have the hyperlinks that arise from the test documents and point to the ones in training, but we use it for only evaluation purposes in the citation recommendation experiments (described below). After stopping and stemming, our vocabulary size is 46,160. The average training full text document is 597.87 words long while the average test abstract is only 45.24 words long.

4.1 Citation Recommendation

This task consists of predicting the true citations (outgoing links) of a document based on its textual content. This evaluation assumes that documents tend to cite other documents that are both topically relevant and topically influential. We believe this assumption holds good in academic literature, and therefore the performance of the TopicFlow model on this task should serve as a good empirical test for the claim that the model captures the notion of topical influence. In our experimental setup, for each abstract in the test set of the ACL corpus, we score documents in the training set based on a “citability score”, as defined by each model. We then rank the predicted citations for each test abstract in the decreasing order of the citability score and evaluate the quality of the ranked list using Mean Average Precision (MAP) measured with respect to its true citations.

As a first baseline citability score, we used TF-IDF based cosine similarity between the test document’s content and the training document’s content. We also ran basic LDA that uses no hyperlink information, and Link-PLSA-LDA [6], a model shown to outperform other baseline joint topic models for text and citations for comparison. For all other topic models we used in our experiments, we computed cosine similarity score between the train and test documents in the corresponding topic space, and combined this score with TF-IDF score using a convex interpolation: score = ζ(Model-score) + (1 − ζ)(TF-IDF-score). For LDA and Link-LDA-PLSA [6], cosine similarity is computed between the test document’s inferred topic distribution and each train document’s learned topic distribution.

For TopicFlow model, we computed cosine between the train document’s topical influence given by \( f_d^{(k)} \) and the test document’s inferred topic distribution \( \theta_d \). Finally, we also used Topic Sensitive PageRank as an additional model for comparison. However, since TSP requires a pre-defined set of topics and a seed set of labeled documents for each topic, we used topics from the TopicFlow model as the TSP topics. In addition, for each topic \( k \), we used all documents \( d \) that satisfy \( \arg \max_k \theta_{dk} = k \) as the seed examples for that topic.

For all the models except the baseline TF-IDF model, we tuned their respective free parameters on a similar time-wise train/test partition of the training set with the number of topics \( K \) fixed at 30. For all models, we found that the best performance is reached in the interval \( \zeta \in [0.05, 0.2] \). For TSP, the optimal value of the teleportation probability is found to be 0.30. For TopicFlow models, we found that tuning the regularization parameter didn’t help, so we fixed it at 0.0. For all models,
we report the MAP scores for $K = 5, 10, 20, 40$ and $60$. The results of our experiments, displayed in Fig. 2, show that TopicFlow model outperforms the strong TF-IDF baseline by as much as 6.4% and is significantly better than the TSP model as per Wilcoxon’s signed rank test at 99% confidence.

4.2 Visualization

Fig. 3 presents a visualization of the TopicFlow model run on the ACL training corpus. Since the corpus consists of papers in the ACL conference, we see mostly Natural Language Processing topics such as “Machine Translation”, “Syntactic Parsing” and “Discourse Analysis”. Note that the top 10 words are highly representative of the topics. As indicated in the bottom row of the table, the model is also able to numerically quantify the influence of documents on each topic. In addition, we also display the flows in the neighborhood of the most influential document on the topic of “Machine Translation”, which helps us understand how influence has spread across the network on this topic. This feature, to the best of our understanding, is very novel and unique to the TopicFlow model.

5 Conclusion

In the future, we plan to compare TopicFlow with the more recent Relational Topic Models [2] on citation recommendation as well as on the more traditional perplexity evaluation. In addition, we plan to build a user-friendly web based graphical browser based on the model’s output on various corpora, that allows us to track the flow of influence of a topic across a large document network.

Acknowledgments

This research was supported by NSF grant NSF-0835614 CDI-Type II: What drives the dynamic creation of science? We wish to thank our anonymous reviewers and the members of the Stanford Mimir Project team for their insights and engagement.

Figure 3: Visualization of a 20 topic TopicFlow model on the ACL training corpus: Left: top 10 words and the most influential document (bottom row) for three representative topics. The numbers in the bottom row in braces indicate the topic-specific influence of the document as measured by $f^{(k)}(d) = f \times 100$. Right: a slice of the TopicFlow model in the neighborhood of the most influential document (bordered in broken lines in dark red) on the topic of "Machine Translation". The numbers next to the arrow are the topic-specific flows, times 100.
References