

Using Text Analysis to Understand the Structure and Dynamics of the World Wide Web as a Multi-Relational Graph

Harish Sethu

Department of ECE, Drexel University
Philadelphia, PA 19104
E-mail: sethu@drexel.edu

Alexander Yates

Department of CIS, Temple University
Philadelphia, PA 19122
E-mail: yates@temple.edu

Abstract—A representation of the World Wide Web as a directed graph, with vertices representing web pages and edges representing hypertext links, underpins the algorithms used by web search engines today. However, this representation involves a key oversimplification of the true complexity of the Web: an edge in the traditional Web graph represents only the existence of a hyperlink; information on the context (e.g., informational, adversarial, commercial, spam) behind the hyperlink is absent. In this *work-in-progress* paper, we describe an ongoing collaborative project between two teams, one specializing in network science and analysis and the other specializing in text analysis and machine learning, to address this oversimplification. Using techniques in natural language processing, text mining and machine learning to extract relevant features of hyperlinks and classify them into one of several types, this undertaking builds and analyzes a multi-relational web graph. A key aspect of this work is that the multi-relational graph emerges naturally from the data instead of being based on an imposed classification of the hyperlinks.

Key words: Web graph; network science; web search; graph sampling; text mining; classification; clustering.

I. INTRODUCTION

The study of the World Wide Web as a directed graph, with vertices representing web pages and edges representing hypertext links between the pages, is central to how content on the Web is discovered, indexed, stored and searched [1]. The properties of the graph are also of scientific interest to sociologists, economists, mathematicians and businesses because of how critical the Web has become to our social, political, commercial and intellectual lives. Research on these properties of the Web graph has taken many different approaches spanning the entire range from theory to experimental measurements. However, almost all of this research uses the traditional graph representation of the Web with a vertex representing each page and a directed edge representing each hyperlink. The existence of an edge carries only an expression of the existence of a hyperlink; information regarding the context (e.g., informational, commercial, friendly, adversarial, spam, organizational) behind the existence of the hyperlink is absent in the graph. This over-simplification in the representation of the Web fails to capture the inherent complexity of the relationships that underlie the Web graph and potentially distorts many of our

conclusions, whether about the importance of pages or about the temporal evolution of the graph.

The variety and complexity of the attributes of hyperlinks do play a role in determining the structure and dynamics of the Web graph and is key to understanding its growth and evolution. Sociologists have already considered multi-relational networks in their research on social networks in recognition of the fact that our social lives are not all based on just one kind of relationship with all others [2]. The Web graph hides a similar layer of complexity in the nature of the hyperlinks. Even though this complexity is recognized in many page ranking algorithms for web search [3], there is no published work on a multi-relational graph representation of the Web based on a methodical categorization of the hyperlinks.

In this work-in-progress paper, we describe a collaborative project between two teams: one specializing in network science and analysis and the other specializing in text analysis and mining. The goal of our work is to develop a systematic characterization of the Web as a multi-relational graph by classifying hyperlinks into one of several types using techniques in natural language processing, text mining and machine learning. The objective is to use text analysis to extract relevant features of hyperlinks that serve as the basis for the classification and allow a multi-relational graph to emerge naturally from the data instead of imposing a classification of the hyperlinks.

We begin with crawling the web to collect a sample of a Web subgraph defined by organizational ownership (e.g., all pages in the `drexel.edu` domain) or by the *forest fire* approach of sampling a graph [4] and catalog the pages and the hyperlinks as a single-relation graph. Similarity measures on the HTML structure of source and target pages and the use of natural language processing tools on the text surrounding the hyperlink in the source document enable us to extract features of hyperlinks relevant to a classification. Using traditional clustering techniques from data mining, such as K-means and hierarchical agglomerative clustering, this work groups web hyperlinks into a small number of categories, each category representing a different type of relationship between the source

and target web pages. The multi-relational graph that emerges is analyzed for its properties relevant to its structure as well as the temporal evolution of the World Wide Web.

In this work-in-progress paper, we report on our methodology and the ongoing effort including some preliminary indications of the promise of our approach.

II. HYPERLINK CLASSIFICATION USING TEXT ANALYSIS

While prior work on hyperlink classification is very limited, a great deal of work has gone into the related problem of determining features for hypertext classification, where the objective is to identify the topic of a web page. The basic feature set for this task is the vector space model of a document’s text. However, studies show that features of a web page’s layout and the hierarchical structure of web directories can improve topic clustering and classification. Several studies have also shown that supplementing a page’s text with anchor text from inbound links, or possibly text from the neighborhood of the anchor, can improve hypertext classification, or even information retrieval [5]–[8]. In particular, as suggested in [5], breaking neighbor pages into segments, or fields, and adding weighted combinations of neighbor pages’ fields into a document representation can significantly help classification. Some studies also suggest that while text from neighboring pages in the Web graph can be helpful, it may hurt if not treated carefully [9], [10]. They also argue that pattern-based extraction over web text can generate features that are sometimes more valuable than the text of a neighbor page. Although this problem is related to link classification, our task on classifying hyperlinks for a multi-relational web graph requires a more careful analysis of the source page of a link, and especially the context around the link anchor, than is typically done for hypertext classification, and our task requires features which represent the relationship between two pages, rather than the topic of a single page.

A. Feature extraction

We begin with annotation techniques to extract distinguishing features of hyperlinks based on the textual content in the source page close to the hyperlink. Besides HTML structure of the source and target pages, the URL and other metadata, the goal of the text analysis is to also consider relevant information for link classification embedded in the text or media content of the page. In the related task of hypertext classification, for example, systems typically use a vector space model of the text in a page, and then add in weighted samples of the text from neighboring pages in the link graph. This approach has two drawbacks from the point of view of link classification: it conflates the text of source and target pages, so that edges from page A to page B are difficult to distinguish from edges from page B to page A . Second, it gives equal weight to all text on a web page, regardless of its position relative to a link, whereas the text immediately surrounding a link is usually the most informative for classifying it.

Let $l \in \mathcal{L}$ be a link consisting of a source HTML document s , a target HTML document t , and an anchor node a in

s . Formally, the goal of our feature extraction task is a representation function $R: \mathcal{L} \rightarrow \mathbb{R}^d$ such that the features $R(l)$ are informative for link classification. In particular, the goal is a representation R that provides features of the local vicinity of a in s and global features of t . Moreover, the features should characterize the relationship between s and t . Two kinds of approaches for characterizing the relationship between the source and target pages of a hyperlink are useful and complementary. One approach measures the similarity between components (anchor text, context window around the text, whole page, metadata) of the source page and target page, and uses each of these similarity scores as features. The other approach builds a vector space model for the words and concepts that characterize the relationship between source and target page. For instance, a vector space model for the intersection of the words on the two pages, or models of the asymmetric set difference between the two pages, can produce a vocabulary for describing how the two pages are related.

In order to make the most of text surrounding a hyperlink anchor, we use a number of annotation tools for natural language processing. Representations that treat documents as bags of words lose a great deal of information. Our approach is to annotate sentences with grammatical and semantic information to recover some of the lost meaning for short segments of text. For instance, a named entity tagger and a semantic role labeling system can identify classes and predicates in text. We can then use named-entity classes and relation types as dimensions in our vector space models, rather than the less-informative word-based representations.

In our early work, we have developed several open-domain NLP pipeline tools for accurately annotating sentences on the Web with various syntactic and semantic information, including part-of-speech tags and chunks, semantic predicates, semantic arguments and roles, and synonymy and paraphrase relationships. Whereas most tools for these tasks perform well on newswire text, the domain that their training data usually comes from, their performance tends to suffer on non-newswire text. Using novel *distributional representations*, and especially representations based on latent variable language models, our tools generalize their classification techniques so that the classifier’s predictions remain accurate on out-of-domain text. These results provide evidence that our systems can port to new domains, and can be quite accurate on web text. The success of our feature extraction techniques may be measured using a variety of standard metrics for measuring the quality of features, including metrics like information gain and measures of the sparsity of the features [11].

B. Hyperlink Classification

We cast the task of identifying relations in a multi-relational network as a classification or clustering problem. Given a set of pages P , a set of hyperlinks L between them, and a representation function R that provides features in \mathbb{R}^d for every hyperlink, our objective is to label each $l \in L$ with a tag that indicates the type of the link.

Our early effort includes the design of a clustering algorithm for sparse, high-dimensional text data that can scale to huge datasets [12]. Our greedy agglomerative clustering technique has been run successfully on a dataset of over 90 million samples taken from a web crawl. It depends on a novel indexing mechanism that drastically cuts down on the number of comparisons made between wholly dissimilar data points. In fact, if we have information regarding the prior distribution of our features, we can provide novel theoretical guarantees on the run-time complexity of our agglomerative clustering. For instance, for special kinds of Zipf distributions (which are common in text data), our clustering algorithm can guarantee that only $O(N \log N)$ comparisons will be made between N samples, as opposed to the usual number of $O(N^2)$ comparisons. In previous work, we have demonstrated that a mutually recursive clustering approach based on this indexing technique can improve the accuracy of the resulting clusters in a synonym detection task [13].

Our work seeks to achieve a meaningful classification in two different ways, arriving at possibly two different results to be iteratively integrated later. Our first approach assumes that we have a fixed set of possible link types, known *a priori*, with examples of each kind of link. We build a system that can classify every link $l \in L$ into one of these types. While this assumption is restrictive, it allows us to develop a framework for evaluating our feature representations and understanding basic aspects of the problem, such as: what kinds of basic link types are most common? And, how does the appearance of one link type on a page affect whether or not a link of another type also appears on that page? In our second approach, we relax the assumption above: we use efficient, large-scale clustering algorithms that, given only the number of link types that appear in L , are able to separate the links of one type from the links of every other type. This allows the multi-relational structure of the network to emerge from the data, rather than being imposed on the network. We expect the results and understanding that we gain from using the first approach to significantly influence our methods and results using the second approach. Integrating the results from these two approaches yields a classification relevant to understanding the structure and dynamics of the web.

III. ANALYZING THE MULTI-RELATIONAL WEB GRAPH

While our work focuses on the World Wide Web, it is but one example of a multi-relational network. In the field of social network analysis, multi-relational networks are most commonly referred to as “multi-modal” [2], [14]. They are usually analyzed as nested networks or networks embedded in networks. However, large data sets on multi-modal social networks (including even online social networks) are not readily available for analysis by academic researchers (because of anti-scraping clauses on social networking web sites) or have not yet been examined in depth. A few works have tried to examine overlapping networks such as the friend networks and the activity networks which form a directed and weighted subgraph of the friend network. There is actually

much evidence that even our offline activities play a role in the online social network and how it grows and develops. As has also been discovered by Kossinets and Watts [15], network evolution is dominated by more than one relationship between the nodes in the network, validating the multi-relational hypothesis of our proposed research. In fact, our own early work on validating this multi-relational nature using an analysis of a million users on YouTube reveals that different types of relationships between users generate networks with important similarities as well as differences in their large-scale properties. For example, independent of the nature of relationships between the users (whether they have provided a video response to, subscribed to, or marked as favorite another user’s videos), the corresponding networks that are generated all exhibit structural properties largely consistent with the power law phenomenon, but yet are substantially different in the degree exponent [16].

A. Structural properties of the Web Graph

The single-relation graph of the Web has been analyzed by a number of different researchers using a variety of different crawling algorithms [17]. Each of these studies has discovered the presence of significant complexity in the topological features of the graph [18], [19]. Among the best-known of these features is the small-world phenomenon (the average number of hyperlinks that one needs to traverse to reach one document from another grows only logarithmically with the growth of the number of URLs) [20], [21]. As in many large networks, the Web graph has also been found to display a power-law relationship between the degrees of nodes and the frequencies of occurrence of nodes of a degree [22]. Our collaborative effort begins with a verification of whether these properties are preserved in the different single-relation networks that make up our multi-relational Web graph. This work reports on additional properties including the distribution of the sizes of the connected components in the graph, in-degree and out-degree correlations at a single node as well as between neighboring nodes, the distribution of sizes of node clusters, and spectral properties of the graph. Ongoing work seeks to further characterize the multi-relational nature of the graph by describing the local, regional (i.e., at the level of neighbors or within a small number of hops) as well as global correlations in structural properties between each of the single-relation networks. Our analytic approach employs soft clustering algorithms based in spectral graph theory, where one can have an affiliation with more than one cluster/group as described in [23].

B. Temporal Evolution of the Web Graph

Toward the goal of understanding the temporal evolution of the web graph, we capture the structural properties of the Web graph at periodic intervals of time and make observations of the temporal patterns as the graph grows and evolves. The selection of the temporal data that we focus on is influenced by the explanatory power of the data and the features that have been of most interest in the science of large-scale dynamics

of networks. For example, proximity bias, reciprocation and preferential attachment are all different ways of understanding the growth of a network and of clusters [24]. In some cases, it is possible to predict whether or not a link will be created between two nodes with an examination of the topology alone [25]. However, in some other cases, as revealed by the adsorption algorithm, proximity alone does not imply shared interest and therefore a cluster. As a result, some authors have tried to detect and predict clusters using a methodology that computes distance measures based on a normalized vector representation of users using network structure as only one of several factors [26]. Temporal growth can also be similarly understood through related methodologies based on latent space models in which each entity is associated with a point in a p -dimensional Euclidean latent space [27]. Our preliminary work indicates that in a complex multi-relational network, one is unlikely to observe only one of these phenomena and it is not even clear if these are the only ones that exist. In this collaborative project, our ongoing work seeks to answer questions on which of these explanations dominate the dynamics of the real Web graph.

IV. CONCLUDING REMARKS

Other than in the field of social network analysis where “multi-modal” graphs have been considered, most research in network science has only considered one kind of relationship between entities in the network. However, relationships that bind together large complex networks are very likely to differ widely in their nature, frequency of use, influence they exert and in their endurance over time. This fact—that large networks in our society are more complex than is observable through the graph generated by any one kind of relationship—is widely recognized but not as often expressed in mathematical treatments in network science. Amongst the same set of entities, different kinds of relationships lead to different sets of edges generating different networks with possibly very different properties.

The nature of the different types of relationships between web pages is largely embedded in the text surrounding the hyperlinks. Text analysis, therefore, unveils a multi-relational perspective on Web graphs contributing a rich new set of data and adding several layers of complexity to the questions one might ask and the answers we might discover about the Web. These answers, we hope, will also help advance the study of multiple co-existing and co-evolving networks other than the Web, including social networks.

REFERENCES

- [1] L. Page, S. Brin, R. Motwani, and T. Winograd, “The pagerank citation ranking: Bringing order to the web,” Stanford InfoLab, Technical Report 1999-66, Nov. 1999. [Online]. Available: <http://ilpubs.stanford.edu:8090/422/>
- [2] R. A. Hanneman and M. Riddle, *Introduction to Social Network Methods*. Riverside, CA, USA: University of California, Riverside, 2005.
- [3] A. Borodin, G. O. Roberts, J. S. Rosenthal, and P. Tsaparas, “Link analysis ranking: algorithms, theory, and experiments,” *ACM Trans. Internet Technol.*, vol. 5, no. 1, pp. 231–297, 2005.
- [4] J. Leskovec and C. Faloutsos, “Sampling from large graphs,” in *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2006, pp. 631–636.
- [5] X. Qi and B. D. Davison, “Classifiers without borders: incorporating fielded text from neighboring web pages,” in *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM, 2008, pp. 643–650.
- [6] E. J. Glover, K. Tsioutsoulouklis, S. Lawrence, D. M. Pennock, and G. W. Flake, “Using web structure for classifying and describing web pages,” in *WWW '02: Proceedings of the 11th international conference on World Wide Web*, 2002, pp. 562–569.
- [7] Y. Yang, S. Slattery, and R. Ghani, “A study of approaches to hypertext categorization,” *J. Intell. Inf. Syst.*, vol. 18, no. 2-3, pp. 219–241, 2002.
- [8] J. Fürnkranz, “Exploiting structural information for text classification on the www,” in *Intelligent Data Analysis*, 1999, pp. 487–498.
- [9] S. Chakrabarti, B. Dom, and P. Indyk, “Enhanced hypertext categorization using hyperlinks,” *SIGMOD Record*, vol. 27, no. 2, pp. 307–318, 1998.
- [10] R. Ghani, S. Slattery, and Y. Yang, “Hypertext categorization using hyperlink patterns and meta data,” in *ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning*, 2001, pp. 178–185.
- [11] A. Dasgupta, P. Drineas, B. Harb, V. Josifovski, and M. W. Mahoney, “Feature selection methods for text classification,” in *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2007, pp. 230–239.
- [12] A. Yates and O. Etzioni, “Unsupervised methods for determining object and relation synonyms on the web,” *Journal of Artificial Intelligence Research (JAIR)*, vol. 34, pp. 255–296, March 2009.
- [13] A. Yates, “Information extraction from the web: Techniques and applications,” Ph.D. dissertation, University of Washington, Department of Computer Science and Engineering, August 2007.
- [14] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*. New York, NY, USA: Cambridge University Press, 1994.
- [15] G. Kossinets and D. J. Watts, “Empirical analysis of an evolving social network,” *Science*, vol. 311, no. 5757, pp. 88–90, January 2006.
- [16] X. Chu and H. Sethu, “Relationships between online social networks and the user-user bonds that create them,” in *Proceedings of the WebSci'09: Society On-Line*, March 2009.
- [17] M. A. Serrano, A. Maguitman, M. Bogu ná, S. Fortunato, and A. Vespignani, “Decoding the structure of the www: A comparative analysis of web crawls,” *ACM Trans. Web*, vol. 1, no. 2, p. 10, 2007.
- [18] S. Dill, R. Kumar, K. S. Mccurley, S. Rajagopalan, D. Sivakumar, and A. Tomkins, “Self-similarity in the web,” *ACM Trans. Internet Technol.*, vol. 2, no. 3, pp. 205–223, 2002.
- [19] D. Donato, L. Laura, S. Leonardi, and S. Millozzi, “The web as a graph: How far we are,” *ACM Trans. Internet Technol.*, vol. 7, no. 1, p. 4, 2007.
- [20] R. Albert, H. Jeong, and A. L. Barabasi, “The diameter of the world wide web,” *Nature*, vol. 401, pp. 130–131, 1999.
- [21] L. Björneborn, “Small-world linkage and co-linkage,” in *HYPERTEXT '01: Proceedings of the 12th ACM conference on Hypertext and Hypermedia*. New York, NY, USA: ACM, 2001, pp. 133–137.
- [22] A. L. Barabási and R. Albert, “Emergence of Scaling in Random Networks,” *Science*, vol. 286, pp. 509–512, October 1999.
- [23] K. Yu, S. Yu, and V. Tresp, “Soft clustering on graphs,” in *Proceedings of the Advances in Neural Information Processing Systems*, 2005.
- [24] A. Mislove, H. S. Koppula, K. P. Gummadi, P. Druschel, and B. Bhat-tacharjee, “Growth of the flickr social network,” in *WOSP '08: Proceedings of the first workshop on Online social networks*. New York, NY, USA: ACM, 2008, pp. 25–30.
- [25] D. Liben-Nowell and J. Kleinberg, “The link prediction problem for social networks,” in *CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management*. New York, NY, USA: ACM, 2003, pp. 556–559.
- [26] M. Maia, J. Almeida, and V. Almeida, “Identifying user behavior in online social networks,” in *SocialNets '08: Proceedings of the 1st workshop on Social network systems*. New York, NY, USA: ACM, 2008, pp. 1–6.
- [27] P. Sarkar and A. W. Moore, “Dynamic social network analysis using latent space models,” *SIGKDD Explor. Newsl.*, vol. 7, no. 2, pp. 31–40, 2005.