Indian J.Sci.Res. 17(2): 596-602, 2018
ISSN: 0976-2876 (Print)
ISSN: 2250-0138(Online)

CONNECTION EXTRACTING: A MINT KNOWLEDGE SEARCH CHALLENGE

[1]Neelam Sriram Kumar , [2]Mallikarajun Kesaratti , [3] K.Naresh
[1,2,3] Department of Computer Science and Engineering, Aurora's Scientific, Technological and Research Academy, Chandrayangutta, Hyderabad.

Abstract - A key challenge for data mining is tackling the problem of mining richly structured datasets, where the objects are linked in some way. Links among the objects may demonstrate cer-tain patterns, which can be helpful for many data mining tasks and are usually hard to capture with traditional statistical mod-els. Recently there has been a surge of interest in this area, fueled largely by interest in web and hypertext mining, but also by interest in mining social networks, security and law enforcement data, bibliographic citations and epidemiological records.
stance of multi-relational data mining (in its broadest sense); however, we use the term link mining to put an additional em-phasis on the links—moving them up to first-class citizens in the data analysis endeavor.
Link mining encompasses a range of tasks including descrip-tive and predictive modeling. Both classification and cluster-ing in linked relational domains require new data mining algo-rithms. But with the introduction of links, new tasks also come to light. Examples include predicting the numbers of links, predicting the type of link between two objects, inferring the existence of a link, inferring the identity of an object, finding co-references, and discovering subgraph patterns. We define these tasks and describe them in more detail in Section 3.

Keywords :

I.Introduction

Traditional data mining tasks such as association rule min-ing, market basket analysis and cluster analysis commonly at-tempt to find patterns in a dataset characterized by a collection of independent instances of a single relation. This is consis-tent with the classical statistical inference problem of trying to identify a model given a random sample from a common underlying distribution.

A key challenge for data mining is tackling the problem of mining richly structured, heterogeneous datasets. These datasets are typically multi-relational; they may be described by a rela-tional database, a semi-structured representations such as XML, or using relational or first-order logic. However, the key com-monalities are that the domain consists of a variety of object types and objects can be linked in some manner. In this case, the instances in our dataset are linked in some way, either by an explicit link, such as a URL, or by a constructed link, such as a join operation between tables stored in a database. Naively applying traditional statistical inference procedures, which assume that instances are independent, can lead to in-appropriate conclusions [24]. Care must be taken that poten-tial correlations due to links are handled appropriately. In fact, record linkage is knowledge that should be exploited. Clearly, this is information that can be used to improve the predictive accuracy of the learned models: attributes of linked objects are often correlated and links are more likely to exist between objects that have some commonality.

Link mining is a newly emerging research area that is at the in-tersection of the work in link analysis [25; 14], hypertext and web mining [3], relational learning and inductive logic pro-gramming [13] and graph mining [8]. Link mining is an in-

II. Background

Probably the most famous example of exploiting link struc-ture is the use of links to improve information retrieval results. Both the well known page rank measure [35] and hubs and au-thority scores [27] are based on the link structure of the web. These algorithms are based on the citation relation between web pages. Recently, many algorithms have been proposed which examine other relations, for example, Dean and Hen-zinger [9] proposed an algorithm based on co-citations to find related web pages, or finer-grained representation of the web pages [5]. Richardson and Domingos [40] combined content and link information with a relevance model to improve per-formance.

A closely related line of work is hypertext and web page clas-sification. This work has its roots in the information retrieval (IR) community. A hypertext collection has a rich structure that should be exploited to improve classification accuracy. In addition to words, hypertext has both incoming and outgoing links. Traditional IR document models do not make full use of the link structure of hypertext. In the web page classifi-cation problem, the web is viewed as a large directed graph. Our objective is to label the category of a web page, based on features of the current page and features of linked neighbors. With the use of linkage information, such as anchor text and neighboring text around each incoming link, better categoriza-tion results can be achieved. Chakrabarti et al. [4] proposed a probabilistic model to utilize both text and

[1]Corresponding Author

linkage information to classify a database of patents and a small web collection. They showed that naively incorporating words from neighbor-ing pages reduces performance, while incorporating category information, such as hierarchical category prefixes, improves performance. Oh et al. [34] reported similar results on a collection of encyclopedia articles: simply incorporating words from neighboring documents was not helpful, while making use of the predicted class of neighboring documents was help-ful. These results indicate that simply assuming that link doc-uments are on the same topic, and incorporating the features of linked neighbors, is not generally effective.

Another approach to hypertext and link mining combines tech-niques from inductive logic programming with statistical learn-ing algorithms to construct features from related documents. A pioneering example is the work of Slattery and Craven [43]. They proposed a model which goes beyond using words in a hypertext document making use of anchor text, neighboring text, capitalized words and alphanumeric words. Using these statistical features and a relational rule learner based on FOIL [39], they proposed a combined model for text classification. Popescul et al. [38] also combined a relational learner with a logistic regression model to improve accuracy for document mining.

Other approaches to link mining identify certain types of hy-pertext regularities such as encyclopedic regularity (in which linked objects typically have the same class) and co-citation regularity (in which linked objects do not share the same class, but objects that are cited by the same object tend to have the same class). Yang et al. [48] gave an in-depth investigation of the validity of these regularities across several datasets and using a range of classifiers. They found that the usefulness of the regularities varied, depending on both the dataset and the classifier being used.

Another link mining task that has received increasing atten-tion is the identification of communities or groups, based on link structure. Gibson et al. [20] gave a survey of work in discovering Web communities. Kubica et al. [29] proposed a probabilistic model for link detection and modeling groups that makes use of demographic information and linkage infor-mation to infer group membership.

Social and collaborative filtering has also been a focus of re-search that can be viewed as link mining. Kautz et al. [26] constructed social networks from Internet data and used the networks to guide users to experts who can answer their ques-tions. Domingos and Richarson [12] modeled the potential value of a customer, based on their network connections.

Others have proposed generative probabilistic models for linked data. Cohn and Hofmann [7] proposed a probabilistic model for hypertext content and links. We also proposed a generative model for relational data, both content and links [17]. How-ever, depending on the task, predictive models may be more appropriate. Examples of predictive modeling in relational do-mains include [44], [38], and [31].

III. Link Mining Tasks

As mentioned in the introduction, link mining puts a new twist on some classic data mining tasks, and also poses new prob-lems. Here we provide a (non-exhaustive) list of possible tasks. We illustrate each of them using the following domains as mo-tivations:

Web page collection: In a web page collection, the objects are web pages, and links are in-links, out-links and co-citation links (two pages that are both linked to by the same page). Attributes include HTML tags, word ap-pearances and anchor text.

Bibliographic domain: In a bibliographic domain, the ob-jects include papers, authors, institutions, journals and conferences. Links include the paper citations, author-ship and co-authorship, affiliations, and the appears-in relation between a paper and a journal or conference.

Epidemiological Studies: In an epidemiology domain, the objects include patients, people they have come in con-tact with, and disease strains. Links represent contacts between people and which disease strain a person is in-fected with.

Link-Based Classification

The most straightforward upgrading of a classic data mining task to linked domains is link-based classification. In link-based classification, we are interested in predicting the cate-gory of an object, based not just on its attributes, but on the links it participates in, and on attributes of objects linked by some path of edges.

An example of link-based classification that has received a fair amount of attention is web-page classification. In this prob-lem, the goal is predict the category of a web page based on words on the page, links between pages, anchor text and other attributes of the pages and the links. In the bibliographic do-main, an example of link-based classification is predicting the category of a paper, based on its citations, the papers that cite it, and co-citations (papers that are cited with this paper). In the epidemiology domain, an example is the task of predicting the disease type based on characteristics of the people (note the arbitrary possible prediction direction) or predicting the person's age, based on the disease they are infected with and the ages of the people they have been in contact with.

Link-based Cluster Analysis

The goal in cluster analysis is to find naturally occurring sub-classes. This is done by segmenting the data into groups, where objects in a group are similar to each other and are very dissimilar from objects in different groups. Unlike classifica-tion, clustering is unsupervised and can be applied to discover hidden patterns from data. This makes it an ideal technique for applications such as scientific data exploration, informa-tion retrieval, computational biology, web log analysis, crimi-nal analysis and many others.

There has been extensive research work on clustering in areas such as pattern recognition, statistics and machine learning. Hierarchical agglomerative clustering (HAC) and k-means are two of the most common clustering algorithms. Probabilistic model-based clustering is gaining increasing popularity [21; 45; 29]. All of these algorithms assume that each object is described by a fixed length attribute-value vector.

In the case of clustering linked data, even the definition of an element in a cluster is open to interpretation. We can clus-ter individual objects, collections of linked objects, or some other subgraph of the original. How do we compare the simi-larity of two of these elements or subgraphs, with potentially different structures? As this may necessitate tests for graph-isomorphism, things will quickly become intractable. There has been surprisingly little work done on this type of link min-ing. Subdue [8] is the earliest line of research in this area. More recent approaches have been focused on efficiently find-ing frequently occurring patterns [30; 23]; these are largely inspired by the apriori algorithm [1] for mining frequently oc-curring patterns. One very interesting new approach is ANF [36], which attempts to compress a graph by approximating the neighborhood function for each node.

Examples of clustering in web page collections range from finding hubs (pages that point to lots of pages of the same cate-gory) to identifying mirror sites. Examples of clustering in the bibliographic domain include finding groups of authors that commonly publish together, and discovering research areas, based on common citations and common publication venues and discovering. An example of clustering in the epidemiol-ogy domain is finding patients with similar sets of contacts or diseases with similar transmission patterns.

Next, we turn to some more specific tasks that arise in link mining. These can often be seen as special cases of link-based classification or link-based cluster analysis.

Identifying Link Type

There is a wide range of tasks related to predicting the exis-tence of links. One of the simplest is predicting the type of link between two entities. For example, we may be trying to predict whether two people who know each other are family members, coworkers, or acquaintances, or whether there is an adviser–advisee relationship between two coauthors.

The link type may be modeled in different ways. In some in-stances, the link type may simply be an attribute of the link. In this case, we may know the existence of a link between two en-tities, and we are simply interested in predicting its type. In our first example, perhaps we know there is some connection be-tween two people, and we must predict whether it is a familial relation, a coworker relation or acquaintance relation. In other instances, there may be different kinds of links. These may be different potential relationships between entities; in the second example, there are two possible relationships: a co-author re-lationship and an adviser–advisee relationship. We may want to make inferences about the existence of one kind of link, having observed another type of link.

A closely related task is predicting the purpose of a link. In a web page collection, the links between pages occur for dif-ferent reasons. At the coarsest grain, links may be for navi-gational purposes or for advertising; it may be quite useful to distinguish between the two. The links may also indicate dif-ferent relationships; the purpose of a link may be to refer to a professor's students, a student's friends, or a course's assign-ments.

Predicting Link Strength

Links may also have weights associated with them. In a web page collection, the weight may be interpreted as the author-itativeness of the incoming link, or its page rank. In an epi-demiological domain, the strength of a link between people may be an indication of the length of their exposure.

Link Cardinality

There are many practical inferences that involve predicting the number of links between objects. The number of links is often a proxy for some more meaningful property whose semantics depend on the particular domain:

In a bibliographic domain, predicting the number of ci-tations of a paper is an indication of the impact of a paper— papers with more citations are more likely to be seminal.

In a web collection, predicting the number of links to a page is an indication of its authoritativeness; predicting the number of links from a page is an indication that the page is a hub. The page rank measure is also clearly related to the number of links.

In an epidemiological setting, predicting the number of links between a patient and people with whom they have been in contact (their contacts) is an indication of the po-tential for disease transmission; predicting the number of links between a particular disease strain and people infected by it is an indication of the strain's virulence.

Note that link counts can be generalized to paths. A count of the number of paths between two objects may be significant.

Record Linkage

Another important concept in link mining is identity uncer-tainty [41; 37; 2]. In many practical problems, such as infor-mation extraction, duplication elimination and citation match-ing, objects may not have unique identifiers. The challenge is to determine when two similar-looking items in fact refer to the same object. This problem has been studied in statistics under the umbrella of record linkage [46; 47]; it has also been studied in the database community for the task of duplicate elimination [42].

In the link mining setting, it is important to take into account not just the similarity of objects based on their attributes, but also based on their links. In the bibliographic setting, this means taking into account the citations of a paper; note that as matches are identified, new matches may become apparent.

IV. Statistical Models For Link Mining

Given the above collection of tasks, there are some unique challenges to applying statical modeling techniques. Here, we identify several; see also other papers in this volume, and pa-pers in several recent workshops on learning statistical models from relational data [18; 19].

Logical vs. Statistical Dependences

The first challenge in link mining and multi-relational data mining is coherently handling two different types of depen-dence structures:

link structure - the logical relationships between ob-jects

probabilistic dependency - the statistical relationship between attributes of objects.

Typically we limit the probabilistic dependence to be among objects that are logically related.

In learning statistical models for multi-relational data, we must not only search over probabilistic dependencies, as is stan-dard in any type of statistical model selection problem, but potentially we must search over the different possible logical relationships between objects. This search over logical rela-tionships has been a focus of research in inductive logic pro-gramming, and the methods and machinery developed in this community should be used to tackle this problem.

Feature Construction

A second challenge is feature construction in the multi-relational setting. The attributes of an object provide a basic description of the object. Traditional classification algorithms are based on these types of object features. In a link-based approach, it may also make sense to use attributes of linked objects. Fur-ther, if the links themselves have attributes, these may also be used. This is the idea behind propositionalization [15; 28]. However, as others have noted, simply flattening the relational neighborhood around an object can be problematic. Several have noted that in hypertext domains, simply including words from neighboring pages degrades classification performance [4; 34]. A further issue is how to deal appropriately with rela-tionships that are not one-to-one. In this case, it may be appro-priate to compute aggregate features over the set of related ob-jects. We have found this works well for learning probabilistic relational models [16], but this approach may not always be appropriate.

Collective Classification

A third challenge is classification using a learned model. A learned link-based model specifies a distribution over link and content attributes, which may be correlated based on the links between them. Intuitively, for linked objects, updating the category of one object can influence our inference about the categories of its linked neighbors. This requires a more com-plex classification algorithm than for a propositional learner. Iterative classification algorithms have been proposed for hy-pertext categorization [4; 34] and for relational learning [33; 45; 44]. The general approach of iterative classification has been studied in numerous fields, including relaxation-labeling in computer vision [22], inference in Markov random fields [6] and loopy belief propagation in Bayesian networks [32]. Some approaches make assumptions about the influence of the neighbor's categories (such as that linked objects have similar categories); we believe it is important to learn how the link dis-tribution affects the category. As an example, this allows us to learn the notion of hubs – e.g., a computer science department homepage is likely to point to a lot of professor homepages.

Effective Use of Unlabeled Data

Recently there has been increased interest in learning using a mix of labeled and unlabeled data. General approaches in-clude semi-supervised learning, co-training and transductive inference. There are some the unique ways in which unla-beled data can be used to improve classification performance in relational domains:

Just as in the case of the classical machine learning frame-work, in which there are no links among the data, unla-beled data can help us learn the distribution over object descriptions.

Links among the unlabeled data (or test set) can provide information that can help with classification.

Links between the labeled training data and unlabeled (test) data induce dependencies that should not be ig-nored.

Link Prediction

A fifth challenge is link discovery, or predicting the existence of links between objects. A range of the tasks that we have described fall under the category of link prediction. A difficulty here is that the prior probability of a link among any set of in-dividuals is typically quite low. While we have had some suc-cess with simple probabilistic models of link existence [17], we believe this is an area where there is much research to be done.

A further challenge is the discovery of common relational pat-terns or subgraphs; some progress has been made in this area [8; 30; 10]; however, this is an inherently difficult problem.

Object Identity

A final challenge is identity detection. How do we infer aliases, i.e., determine that two objects refer to the same individual? As mentioned earlier, some work has been done in this area by several research communities, but there is a great deal of room for additional work.

Another aspect of this challenge is whether our statistical mod-els refer explicitly to individuals, or only to classes or cate-gories of objects. In many cases, we'd like to model that a connection to a particular object or individual is highly pre-dictive; on the other hand, if we'd like to have our models generalize and be applicable to new, unseen objects, we also have to be able to model with and reason about generic collec-tions of objects.

V. Conclusion

There has been a growing interest in learning from linked data, which are described by a graph in which the nodes in the graph are objects and the edges/hyper-edges in the graph are links— or relations—between objects. Tasks include hypertext classi-fication, segmentation, information extraction, searching and information retrieval, discovery of authorities and link discov-ery. Domains include the world-wide web, bibliographic ci-tations, criminology and bio-informatics, to name just a few. Learning tasks range from predictive tasks, such as classifica-tion, to descriptive tasks, such as the discovery of frequently occurring sub-patterns. We have given a brief summary of some of the work in this area, and some of the challenges in link mining. Link mining is a promising new area where re-lational learning meets statistical modeling; we believe many new and interesting machine learning research problems lie at the intersection, and it is a research area “whose time has come” [11].

VI. Acknowledgments

This study was supported in part by the Advanced Research

and Development Activity (ARDA) under Award Number NMA401-02-1-2018. The views, opinions, and findings contained in this report are those of the author(s) and should not be construed as

an official Department of Defense position, policy, or decision unless so designated by other official documentation.

References

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In J. B. Bocca, M. Jarke, and C. Zan-iolo, editors, Proc. 20th Int. Conf. Very Large Data Bases, VLDB, pages 487–499. Morgan Kaufmann, 12– 15 1994.

[2] M. Bilenko and R. J. Mooney. On evaluation and training-set construction for duplicate detection. under review.

[3] S. Chakrabarti. Mining the Web. Morgan Kaufman, 2002.

[4] S. Chakrabarti, B. Dom, and P. Indyk. Enhanced hy-pertext categorization using hyperlinks. In Proc of SIGMOD-98, 1998.

[5] S. Chakrabarti, M. Joshi, and V. Tawde. Enhanced topic distillation using text, markup tags, and hyperlinks. In Research and Development in Information Retrieval, pages 208–216, 2001.

[6] R. Chellappa and A. Jain. Markov random fields: theory and applications. Academic Press, Boston, 1993.

[7] D. Cohn and T. Hofmann. The missing link - a proba-bilistic model of document content and hypertext con-nectivity. In Neural Information Processing Systems 13, 2001.

[8] D. Cook and L. Holder. Graph-based data mining. IEEE Intelligent Systems, 15(2):32–41, 2000.

[9] J. Dean and M. R. Henzinger. Finding related pages in the World Wide Web. Computer Networks, 31(11– 16):1467–1479, 1999.

[10] L. Dehaspe, H. Toivonen, and R. D. King. Find-ing frequent substructures in chemical compounds. In R. Agrawal, P. Stolorz, and G. Piatetsky-Shapiro, editors, 4th International Conference on Knowledge Discovery and Data Mining, pages 30–36. AAAI Press., 1998.

[11] P. Domingos. Prospects and challenges for multi-relational data mining. SIGKDD Explorations, 2003. In this volume.

[12] P. Domingos and M. Richardson. Mining the network value of customers. In Proceedings of the Seventh Inter-national Conference on Knowledge Discovery and Data Mining, 2001.

[13] S. Dzeroski and N. Lavrac, editors. Relational Data

Min-ing. Kluwer, Berlin, 2001.

[14] R. Feldman. Link analysis: Current state of the art. In KDD-02 Tutorial, 2002.

[15] P. A. Flach and N. Lavrac. The role of feature construc-tion in inductive rule learning. In Proc. of the ICML2000 workshop on Attribute-Value and Relational Learning: crossing the boundaries, 2000.

[16] L. Getoor, N. Friedman, D. Koller, and A. Pfeffer. Learn-ing probabilistic relational models. In S. Dzeroski and N. Lavrac, editors, Relational Data Mining, pages 307– 335. Kluwer, 2001.

[17] L. Getoor, N. Friedman, D. Koller, and B. Taskar. Learn-ing probabilistic models with link uncertainty. Journal of Machine Learning Research, 2002.

[18] L. Getoor and D. Jensen. Proc. AAAI-2000 Workshop on Learning Statistical Models from Relational Data. AAAI Press, 2000.

[19] L. Getoor and D. Jensen. Proc. IJCAI-2003 Workshop on Learning Statistical Models from Relational Data. AAAI Press, 2003.

[20] D. Gibson, J. M. Kleinberg, and P. Raghavan. Inferring web communities from link topology. In UK Conference on Hypertext, pages 225–234, 1998.

[21] T. Hofmann. Probabilistic Latent Semantic Indexing. In Proceedings of the 22nd Annual ACM Conference on Re-search and Development in Information Retrieval, pages 50–57, Berkeley, California, August 1999.

[22] R. Hummel and S. Zucker. On the foundations of re-laxation labeling processes. IEEE Transactions on Pat-tern Analysis and Machine Intelligence, 39(5):267–287, 1983.

[23] A. Inokuchi, T. Washio, and H. Motoda. An apriori-based algorithm for mining frequent substructures from graph data. In Principles of Data Mining and Knowledge Discovery, pages 13–23, 2000.

[24] D. Jensen. Statistical challenges to inductive inference in linked data. In Seventh International Workshop on Artifi-cial Intelligence and Statistics, 1999.

[25] D. Jensen and H. Goldberg. AAAI Fall Symposium on AI and Link Analysis. AAAI Press, 1998.

[26] H. Kautz, B. Selman, and M. Shah. Referral Web: Com-bining social networks and collaborative filtering. Com-munications of the ACM, 40(3):63–65, 1997.

[27] J. Kleinberg. Authoritative sources in a hyperlinked en-vironment. Journal of the ACM, 46(5):604–632, 1999. 0

[28] S. Kramer, N. Lavrac, and P. Flach. Propositionalization approaches to relational data mining. In S. Dzeroski and N. Lavrac, editors, Relational Data Mining, pages 262— 291. Kluwer, 2001.

[29] J. Kubica, A. Moore, J. Schneider, and Y. Yang. Stochas-tic link and group detection. In Proc. of AAAI-02, 2002.

[30] M. Kuramochi and G. Karypis. Frequent subgraph dis-covery. In ICDM, pages 313–320, 2001.

[31] Q. Lu and L. Getoor. Link-based classification. In Proc. of ICML-03, 2003.

[32] K. Murphy and Y. Weiss. Loopy belief propagation for approximate inference: an empirical study. In Proc. of UAI-99. Morgan Kaufman, 1999.

[33] J. Neville and D. Jensen. Iterative classification in rela-tional data. In Proc. AAAI-2000 Workshop on Learning Statistical Models from Relational Data. AAAI Press, 2000.

[34] H.-J. Oh, S. H. Myaeng, and M.-H. Lee. A practical hy-pertext categorization method using links and incremen-tally available class information. In Proc. of SIGIR-00, 2000.

[35] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bring order to the web. Tech-nical report, Stanford University, 1998.

[36] C. Palmer, P. Gibbons, and C. Faloutsos. Anf: A fast and scalable toole for data mining in massive graphs. In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002), 2002.

[37] H. Pasula, B. Marthi, B. Milch, S. Russell, and I. Sh-pitser. Identity uncertainty and citation matching. In Ad-vances in Neural Information Processing Systems 15 (NIPS 2002). MIT Press, 2003.

[38] A. Popescul, L. Ungar, S. Lawrence, and D. Pennock. Towards structural logistic regression: Combing rela-tional and statistical learning. In KDD Workshop on Multi-Relational Data Mining, 2002.

[39] J. R. Quinlan and R. M. Cameron-Jones. FOIL: A midterm report. In Proc. of ECML-93, 1993.

[40] M. Richardson and P. Domingos. The Intelligent Surfer: Probabilistic Combination of Link and Content Informa-tion in PageRank. In Advances in Neural Information Processing Systems 14. MIT

Press, 2002.

[41] S. Russell. Identity uncertainty. In Proc. of IFSA-01, Vancouver, 2001.

[42] S. Sarawagi and A. Bhamidipaty. Interactive deduplica-tion using active learning. In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002), 2002.

[43] S. Slattery and M. Craven. Combining statistical and re-lational methods for learning in hypertext domains. In Proc. of ILP-98, 1998.

[44] B. Taskar, P. Abbeel, and D. Koller. Discriminative prob-abilistic models for relational data. In Proc. of UAI-02, pages 485–492, Edmonton, Canada, 2002.

[45] B. Taskar, E. Segal, and D. Koller. Probabilistic classifi-cation and clustering in relational data. In Proc. of IJCAI-01, 2001.

[46] W. E. Winkler. Advanced methods for record linkage. Technical report, Statistical Research Division, U.S. Census Bureau, 1994.

[47] W. E. Winkler. Methods for record linkage and bayesian networks. Technical report, Statistical Research Divi-sion, U.S. Census Bureau, 1994.

[48] Y. Yang, S. Slattery, and R. Ghani. A study of approaches to hypertext categorization. Journal of Intelligent Infor-mation Systems, 18(2-3):219–241, 2002.