A Model of Uncertainty for Near-Duplicates in Document Reference Networks

Claudia Hess[1] and Michel de Rougemont[2]

[1] Laboratory for Semantic Information Technology, Bamberg University
[2] LRI, Université Paris-Sud 11
claudia.hess@wiai.uni-bamberg.de, mdr@lri.fr

Abstract. We introduce a model of uncertainty where documents are not uniquely identified in a reference network, and some links may be incorrect. It generalizes the probabilistic approach on databases to graphs, and defines subgraphs with a probability distribution. The answer to a relational query is a distribution of documents, and we study how to approximate the ranking of the most likely documents and quantify the quality of the approximation. The answer to a function query is a distribution of values and we consider the size of the interval of Minimum and Maximum values as a measure for the precision of the answer.

1 Introduction

Digital libraries often contain duplicates, i.e., two or more representations of the same or nearly the same document. Duplicates are, for example, the pre-print and print, or erroneous copies of a document as in the metadata provided by CiteSeer[3]: only around 500,000 of the over 700,000 documents have a distinct title (almost identical titles are hereby not yet filtered). The fraction of duplicated pages on the web was estimated at 30 to 45% in [1, 2]. Duplicates may be mirrors, but also be malicious copies by spammers, or crawling errors. When two heterogeneous document repositories are integrated, the merged collection may contain duplicates, too. Cleaning mechanisms try to avoid duplicates and hence define some identity between objects (e.g. [3]). Objects are merged if their similarity is above a certain threshold. However, a merge is not appropriate when documents differ too much with respect to their metadata, references or content.

We consider a Document Reference Network as a graph where nodes are documents and edges link one document with its references. PageRank [4] is the best known measure which analyzes document reference networks in order to rank the results of user queries. Measures on a document network may provide misleading results if the network contains duplicates. While a duplicate's citation list might be incomplete because not all references were correctly extracted, incoming references might point only to one of the duplicates. This could wrongly increase or decrease the rank of a document. We therefore propose a model of uncertainty for near duplicates. It follows the approach taken by probabilistic

[1] The work was supported by the German Academic Exchange Service.
[3] The CiteSeer metadata is publicly available at http://citeseer.ist.psu.edu/oai.html.

databases where alternative representations of the same real-world entity are available. As several nodes may represent the same document and link to different documents, one may first cluster similar documents and introduce probabilistic edges between the clusters. This simple model captures some of the difficulties to approximate queries in a digital library, and provides some measures of quality for the answers of relational and functional queries. The answer to a unary query is a distribution on documents, and we study how to approximate the sequence of most likely documents and propose a measure for the quality of the approximation. The answer to a functional query is a probabilistic distribution of values from a Min value to a Max value. The smaller is the interval [Min, Max], the better is the precision of the answer. To this end, the paper is structured as follows: section 2 presents the uncertainty model for graphs. Section 3 defines queries in this model and the main section 4 presents an efficient approximation of these queries.

2 Uncertainty Model for Document Networks

We extend the uncertainty model by Andritsos et al. [5] for relational databases to an uncertainty model for graphs. Table 1 shows an example database and figure 1 the corresponding graph with clusters. A cluster-based uncertainty graph is a structure $GC_n = (D_n, E, U_1, ...U_p, C_1, ...C_m)$ where D_n is the set of n nodes, $E \subseteq D_n \times D_n$ is the set of edges, $U_i \subseteq D_n$ is the set of labels, for $i = 1, ..., p$, and C_i are partial functions from D_n into $[0, 1]$ such that the domains of C_i partition D_n, and $\sum_x C_i(x) = 1$.

	id	docID	references	prob
t1	d1	doc1	doc3, doc5	0.7
t2	d1	doc2	doc5	0.3
t3	d2	doc3	doc5, doc6	0.2
t4	d2	doc4	doc5	0.8
t5	d3	doc5		1
t6	d4	doc6		1

Table 1. Document Relation

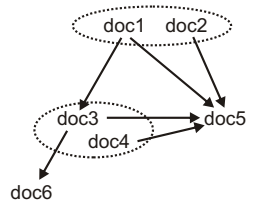

Fig. 1. Graph with Clusters

According to [5], probabilistic instances $\widehat{G}_i$ take one tuple t out of each cluster C_j from G with probability $C_j(t) = prob(t)$. Then $Prob(\widehat{G}_i) = \prod_{t \in \widehat{G}_i} prob(t)$ gives the probability distribution over all $\widehat{G}_i$. The answer to a query Q is defined as a probabilistic measure on tuples: $p_t = \sum_{\widehat{G}_i | t \in Q(\widehat{G}_i)} Prob(\widehat{G}_i)$. Our extension of the relational model indicates how to set edges from and to uncertain nodes in the probabilistic instances: there is an edge between C_i and C_j if the selected nodes $c \in C_i$ and $c' \in C_j$ were connected in GC_n.

We transform the graph with clusters on the nodes GC_n into a graph $G = (V, E_C, \mu_e)$ where V is the set of clusters. The probability μ_e of an edge $e \in E_C$ between two clusters C_i and C_j is the probability over the choices over $c \in C_i$ and $c' \in C_j$ that $(c, c') \in E$.

3 Queries to Uncertain Graphs

Definition 1. *A relational query of arity m on a graph G is a function $Q : G \times V^k \rightarrow R$ where $R \subseteq V^m$.*

In the cluster-based model, each instance $\widehat{G_i}$ provides a random variable $\widehat{R_i}$ and gives rise to a distribution $\mathcal{R} = \{(t, p_t)\}$ of tuples t with probabilities p_t, where $t \in R$ and $p_t = \sum_i \text{prob}(\widehat{G}_i)$ for i such that $\widehat{G}_i \models \widehat{R}(t)$.

Definition 2. *A function query f of arity k on a graph G is: $f : G \times V^k \rightarrow \mathbf{R}$.*

Each instance $\widehat{G}_i$ provides a function $\widehat{f_{\widehat{G}_i}}$ for f, giving a distribution of values (t, p_t) where $t \in \mathbf{R}$ and probability $p_t = \sum_i \text{prob}(\widehat{G}_i)$ for i such that $\widehat{G}_i \models \widehat{f_{\widehat{G}_i}} = t$. The expected function is defined as $E(f) = \sum_{i=1}^{m} \text{prob}(\widehat{G}_i) \cdot \widehat{f_{\widehat{G}_i}}$. We approximate the interval $I = [\alpha, \beta]$, where $\alpha = Min_{\widehat{G}_i} \widehat{f_{\widehat{G}_i}}$ and $\beta = Max_{\widehat{G}_i} \widehat{f_{\widehat{G}_i}}$.

4 Approximation

4.1 Approximation of Relational Queries

A unary relation query Q defines a distribution $\mathcal{R}$ on documents, and a sequence $s = d_{i_1}, d_{i_2}, ..., d_{i_n}$ ordered by decreasing probability. We want to approximate the k first answers, i.e. produce a sequence $s_k = d_{i'_1}, d_{i'_2}, ..., d_{i'_k}$ close to s. A classical distance between two sequences is the *Kendall Tau distance* which measures the number of misclassified pairs (see e.g. [6]). We relativize the weight of a misclassified pair with the difference of their probabilities. For each d in s_k, d' in s is *misclassified (mis.)* for d if d' is not a prefix of d in s_k and $d' > d$ in s.

Definition 3. *The peudo-distance between s_k and the sequence s associated with a distribution $\mathcal{R}$ is:*

$$d(s_k, s) = \frac{\sum_{d \in s_k} \sum_{d' \in s \ mis.} |Prob(d') - Prob(d)|}{N_k}$$

with $N_k = n - 1 + n - 2 + ...n - k$ the maximal number of misclassified pairs.

Definition 4. *A randomized algorithm $\mathcal{A}(G_n, Q, k)$ which outputs a sequence s_k, ϵ-approximates the answer $\mathcal{R}$ if s_k is ϵ-close to the sequence s, with high probability.*

Naive Sampling algorithm. Take N samples $\widehat{G}_i$, evaluate Q and obtain $\widehat{R_i}$. Let c the function which associates with a document d, the number of occurrences of d in $\widehat{R_1}...\widehat{R_N}$. Rank the documents according to c and select the k first answers.

Theorem 1. *The Naive Sampling algorithm ϵ-approximates any unary Least-Fixed point query Q in polynomial time.*

We now quantify the quality of the answer. Consider $s_k = (d_{i_1}, ...d_{i_k})$, where by definition $c(d_{i_j}) \geq c(d_{i_{j+1}})$ for $1 \leq j < k$ and each $c(d_{i_j}) \leq N$. The *quality* of s_k is $\sum c(d_{i_j})/N^2$ which is 1 if all k documents in s_k are present in the answers of all samples, and $1/N$ if each document is only present in one sample.

4.2 Approximation of Functional Queries

We show as an example the approximation of the length of the shortest path between two nodes, which is a basic function on graphs. Measures such as PageRank can be approximated in the same style. The approximation uses the graph G and conditional probabilities on the edges.

Definition 5. *An algorithm $\mathcal{A}$ which outputs $(\alpha,\ \beta)$ ϵ-approximates f if:* $(a)\ f(\widehat{G}_i, u, v) \in [\alpha - \epsilon,\ \beta + \epsilon]$ *for all* $\widehat{G}_i$, $(b)\ \alpha - \epsilon \leq \min_{\widehat{G}_i}\ f(\widehat{G}_i, u, v) \leq \alpha + \epsilon$ *and* $(c)\ \beta - \epsilon \leq \max_{\widehat{G}_i}\ f(\widehat{G}_i, u, v) \leq \beta + \epsilon$.

Shortest Path Approximation. We approximate the shortest path $SP(d_s, d_t)$ from a node d_s to a target d_t in GC_n. We aim to give an interval $I_{d_s \to d_t}$ such that $\widehat{\mathrm{SP}}(d_s, d_t) \in I_{d_s \to d_t} = [\alpha,\ \beta]$. SP is approximated by forwarding intervals in G from d_s to d_t in a naive way.

Interval Propagation Algorithm(GC_n, d_s, d_t)**.** Compute the intervals $I_{d_s \to d_i}$ for d_i connected at distance 1,2,...i from d_s until d_t is reached or all nodes of the connected components C of d_s are reached. If $d_t \notin C$ then $I_{d_s \to d_t} = [\infty, \infty]$. By induction on the depth i we can prove that $\mathcal{A}$ satisfies the properties (a),(b),(c).

Theorem 2. *For each node d_i at depth i from d_s, $\mathcal{A}$ approximates SP with $\epsilon = 0$, after exploring at most n nodes.*

5 Conclusion

Most approaches to query answering over document networks assume networks without duplicates or incorrect links. However, these duplicates distort the results. We developed a model of cluster-based uncertainty for graphs. The answer to unary relational queries is a distribution of documents, and the answer to functional queries is a probabilistic distribution of values ranging from a Min to a Max value. We efficiently approximated these distributions and provided a quality measure for the answers.

References

1. Broder, A.Z., Glassman, S.C., Manasse, M.S., Zweig, G.: Syntactic clustering of the web. Computer Networks **29**(8-13) (1997) 1157–1166
2. Shivakumar, N., Garcia-Molina, H.: Finding near-replicas of documents on the web. In: Proceedings of Workshop on Web Databases (WebDB'98). (1998)
3. Broder, A.Z.: Identifying and filtering near-duplicate documents. In Giancarlo, R., Sankoff, D., eds.: Proceedings of the 11th Annual Symposium on Combinatorial Pattern Matching. Volume 1848 of LNCS., Springer-Verlag (2000) 1–10
4. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project (1998)
5. Andritsos, P., Fuxman, A., Miller, R.J.: Clean answers over dirty databases: A probabilistic approach. In: Proceedings of the International Conference on Data Engineering (ICDE). (2006)
6. Fagin, R., Kumar, R., Mahdian, M., Sivakumar, D., Vee, E.: Comparing and aggregating rankings with ties. In: ACM Principles on Databases Systems, ACM Press (2004) 47–58