Characterization of a simple case of the reassignment of document identifiers as a pattern sequencing problem

Roi Blanco
University of A Coruña
Computer Science Department
Spain
rblanco@udc.es

Alvaro Barreiro
University of A Coruña
Computer Science Department
Spain
barreiro@udc.es

ABSTRACT

In this poster, we analyze recent work in the document identifiers reassignment problem. After that, we present a formalization of a simple case of the problem as a PSP (Pattern Sequencing Problem). This may facilitate future work as it opens a new research line to solve the general problem.

Categories and Subject Descriptors: H.3 [Information Storage And Retrieval]

General Terms: Efficiency, Compression

Keywords: Inverted Files, Compression, Document Identifier Reassignment

1. INTRODUCTION AND MOTIVATION

The reassignment of document identifiers is a recent technique to enhance static compression in inverted files. Some works demonstrated that it is possible to lower the number of bits required to code each posting list by reassigning the document identifiers of the collection. Concretely, for a text collection of N documents and T terms, an inverted file stores a set of T posting lists following the format:

$$< t_i; f_{t_i}; d_{i1}, d_{i2}, \ldots, d_{if_{t_i}} >, d_{ik} < d_{ij} \forall k < j \qquad (1)$$

where f_{t_i} stands for the frequency of the term t_i (number of documents in which t_i appears), and d_{ik} is the k-th document identifier for the term i. As the notation implies, the document identifiers are ordered. A general way of compressing posting lists is to code differences between document identifiers, $d_{ik+1} - d_{ik}$, called d-gaps [6].

The document reassignment problem tries to find the bijective function f that maps each document identifier into a new identifier in the range $[1 \ldots N]$ and minimizes the cost of coding the document gaps. Static codes exploit the fact that small document differences occurs often, assigning shorter codes to them. Indeed, it is clear that reordering the document identifiers in such a way that lower differences between document identifiers occur in all or most posting lists, the resulting total number of bits will also be reduced.

Research in the document identifier reassignment problem is very recent. The first solutions of the problem are presented in [2] and [4], in the following of this paper, the *B&B* (Blandford and Blelloch) and the *TSP* approach (Traveling Salesman Problem) respectively. Both solutions build a *weighted similarity graph* G where the nodes v_i, v_j represent the document identifiers i, j and an edge $e(v_i, v_j)$ represents the similarity between documents i and j.

The *B&B* algorithm recursively splits G into small subgraphs $G_{l,i} = (V_{l,i}, E_{l,i})$ representing smaller subsets of the collection until every subgraph becomes a singleton. After that, the technique performs a reordering of the document identifiers, by *depth-first* traversal. The TSP-based solutions consider the problem a *Traveling Salesman Problem* which can be solved by several heuristics identified in graph theory literature. This technique tries to find the traverse that minimizes the sum of distances between consecutive documents. The minimal traversal gives the new order for the document identifiers. The work in [1] shows that an efficient implementation of the TSP approach is feasible by using a prior dimension reduction with SVD (Singular Value Decomposition).

Although presenting good results in d-gap reduction and therefore in compression ratio gains, the graph-based approaches proposed so far, show scalability problems in terms of space and time. On the other hand, the work in [5] proposes a different approach by *assigning* the document identifiers *on the fly* during the inversion of the text collection. This approach performs well, but under the assumption that the average document length is small.

Given the limitations of previous works, we present a formalization of the problem of minimizing the average d-gap as a pattern sequencing problem (PSP [3]). This minimization problem implies obtaining a solution for the case of coding the d-gaps with unary code. Although this encoding has no practical consequences, the NP-Completeness of the PSP has been proved in the literature [3] and our work opens a research line to address the general problem.

2. FORMALIZATION

Consider a binary matrix C where the columns are document vectors and the rows represent the presence/absence of each term in every document. Working with this binary matrix, the problem of minimizing the average d-gaps consists of finding the permutation of the matrix columns that minimizes a function cost ϕ that computes the total d-gap length.

First, let us consider the posting list format in Equation (1), where for each term t_i the ordered identifiers fall in the range $d_{i1} \ldots d_{if_{t_i}}$. The cost function that measures the

SIGIR'05, August 15–19, 2005, Salvador, Brazil.
ACM 1-59593-034-5/05/0008.

average d-gap, without including the first offset is

$$\phi = \sum_{k=1}^{T} \frac{1}{f_{t_k} - 1} \sum_{i=2}^{f_{t_k}} d_{ki} - d_{k(i-1)} =$$

$$\sum_{k=1}^{T} \frac{1}{f_{t_k} - 1} \left(d_{kf_{t_k}} - d_{k1} \right) \tag{2}$$

In this situation, given a permutation π of the matrix C, C_π, the cost function ϕ is the difference between the column number for the last and first one in each row, divided by the total number of ones minus one (in the row):

$$\phi(\pi) = \sum_{k=1}^{T} \frac{1}{f_{t_k} - 1} \left(\mathbf{max}\{j|_{c_{\pi_{j,k}}>0}\} - \mathbf{min}\{j|_{c_{\pi_{j,k}}>0}\} \right) =$$

$$\sum_{k=1}^{T} \frac{1}{f_{t_k} - 1} \sum_{i=\mathbf{min}\{j|_{c_{\pi_{j,k}}>0}\}}^{\mathbf{max}\{j|_{c_{\pi_{j,k}}>0}\}-1} 1 =$$

$$\sum_{k=1}^{T} \frac{1}{\gamma_k} \left[\sum_{i=\mathbf{min}\{j|_{c_{\pi_{j,k}}>0}\}}^{\mathbf{max}\{j|_{c_{\pi_{j,k}}>0}\}} \alpha_{\pi_i} \right] - \sum_{k=1}^{T} \frac{1}{\gamma_k}, \tag{3}$$

if γ_k stands for the inverse term frequencies minus one, and $\alpha_{\pi_i} = 1, \forall i : 1 \leq i \leq N$.

In the last form of $\phi(\pi)$ in 3, the first term is the expression of the function cost in the Actor Costs (PSP-AC) of the shooting schedules problem [3]. The AC is a generalization of the Average Order Spread (PSP-AOS) problem [3], and both are pattern sequencing problems. In these problems the goal is to find a permutation of predetermined production patterns. We also could consider the document identifier reassignment problem as a PSP-AOS in the particular case that we measure the global average d-gap instead of the per posting list average d-gap. In [3] there is a proof that the PSP-AOS is a NP-complete problem. Therefore, the AC problem is also NP-complete and so this case of the document identifier reassignment.

Actually, coding a posting list implies taking into account the first document identifier appearing in the sequence. We can include this offset into the equation by simply adding it into the sum. Therefore, considering the first offset in each posting list, the cost function is

$$\phi' = \sum_{k=1}^{T} \frac{1}{f_{t_k}} d_{kf_{t_k}} \tag{4}$$

The solution to the problem formalized here is also a solution to the reassignment of document identifiers using unary codes for d-gap encoding. However, in the real case the minimization must take into account the coding scheme. Most static codes represent an integer x with $O(log(x))$ bits. In order to obtain good compression ratios, a better goal would be to minimize the d-gap products.This way, the log of the products and the sum of the logs would be minimal, with practical consequences in the final coded posting lists.

It is important to notice that considering the problem a TSP, although producing good results, is only a strategy to address the document identifier reassignment problem. Given the weighted similarity graph G, the TSP finds the traversal $\langle v_1, \ldots, v_n \rangle$ which maximizes $\sum_{i=1}^{N-1} e(v_i, v_{i+1})$. In fact, the exact TSP solution may not optimize the average sum of d-gaps, as is illustrated in the following example, where the similarity between documents is measured with the inner product. Consider the following 4×4 matrix:

$$\begin{array}{c} \begin{array}{cccc} d_1 & d_2 & d_3 & d_4 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \end{pmatrix} \end{array}$$

The traversal $tr_1 = \langle d_1, d_4, d_2, d_3 \rangle$ is a solution to the TSP as it maximizes $\sum_{i=1}^{N-1} e(v_i, v_{i+1})$, obtaining a value of 4.

$$\begin{array}{c} \begin{array}{cccc} d_1 & d_4 & d_2 & d_3 \end{array} \\ \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix} \end{array}$$

However, the traversal $tr_2 = \langle d_2, d_4, d_1, d_3 \rangle$, which is not a TSP solution, has a lower d-gap sum. For the traversal tr_1 the sum of d-gaps is 12, and for tr_2 is 10.

$$\begin{array}{c} \begin{array}{cccc} d_2 & d_4 & d_1 & d_3 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix} \end{array}$$

Please notice that for this example, there exists another solution for the TSP with traversal $tr_3 = \langle d_3, d_2, d_4, d_1 \rangle$, also with a value of 12 for the sum of d-gaps. Also, the reader can verify that in these examples the TSP solutions are not optimal with respect to the sum of the d-gaps logs, since again the traversal tr_2 is better than tr_1 and tr_3 with respect to this cost function.

3. CONCLUSIONS

We have formalized a particular case of the document identifier reassignment problem, defining the real cost function and identifying the problem as a PSP. This formalization shows the NP-Completeness of the problem. We have proved that TSP solutions can be not optimal, thus leading to the need of experimenting with real-cost d-gap based heuristics.

Acknowledgements: The work reported here was co-founded by the "Secretaría de Estado de Universidades e Investigación" and FEDER funds under research projects TIC2002-00947 and Xunta de Galicia under project PGIDT03PXIC10501PN.

4. REFERENCES

[1] R. Blanco, A. Barreiro. Document identifier reassignment through dimensionality reduction. *Proceedings of the 27th European Conference on Information Retrieval, ECIR 2005*, LNCS 3408, pp. 375-387, 2005.

[2] D. Blandford and G.Blelloch. Index compression through document reordering. *Proceedings of the IEEE Data Compression Conference (DCC'02)*, pp. 342-351, 2002.

[3] A. Fink. S. Voß. Applications of modern heuristic search methods to pattern sequencing problems. *Computers & Operations Research*, 26:17-34, 1999.

[4] W.-Y. Shieh, T.-F. Chen, J. J.-J. Shann and C.-P. Chung. Inverted file compression through document identifier reassignment. *Information Processing and Management*, 39(1):117-131, January 2003.

[5] F. Silvestri, S. Orlando and R. Perego. Assigning identifiers to documents to enhance the clustering property of fulltext indexes. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 305-312, 2004.

[6] I. H. Witten, A. Moffat and T. C. Bell. *Managing Gigabytes - Compressing and Indexing Documents and Images*, 2nd edition. Morgan Kaufmann Publishing, San Francisco, 1999.