Characterization of a Simple Case of the Reassignment of Document Identifiers

Roi Blanco and Alvaro Barreiro
Computer Science Department
University of A Coruña Spain
{rblanco,barreiro}@udc.es

Inverted Files

- Given a text collection of N documents and T terms, an inverted file is a set of posting lists, storing information for every $t_i \in T$
- Each posting list follows the form

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

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.

Inverted File

Real Inverted File

- Actually, the inverted file stores the difference between two consecutive identifiers.
- For each term t_i, the posting list would be $d_{i1}, d_{i2} - d_{i1}, d_{i3} - d_{i2}, \ldots$

D-gaps

- This fact is used by static codes to improve compression.

Reassignment of Doc. Ids

- The objective of the problem is to find a bijective function

$$\begin{array}{c} f : [1..N] \rightsquigarrow [1..N] \\ d_{ij} \longrightarrow \ d'_{ij} \end{array}$$

- In the example: $1 \rightarrow 2; 2 \rightarrow 5; 3 \rightarrow 6$ $4 \rightarrow 7; 5 \rightarrow 3; 6 \rightarrow 4; 7 \rightarrow 1$

D-gaps after reordering

- . . . trying to minimize the cost of coding the doc. id. differences:

$$\phi = \sum_{i=1}^{T} \left[s(d_{i1}) + \sum_{k=2}^{f_{t_i}} s(d_{ik} - d_{i(k-1)}) \right]$$

where $s(x)$ is the number of bits used to code an integer x.

Background

- Previous works approached the problem with heuristic solutions, without considering the real cost function.
- Cluster Based: Blandford and Blelloch IEEE DCC'02 [2], Silvestri et al. SIGIR'04 [5].
- TSP-Based: Shie et al. IP&M'03 [4], Blanco and Barreiro ECIR'05 [1].
  - These 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 goal is to find the traverse that minimizes the sum of distances between consecutive documents. The minimal traversal gives the new order for the document identifiers.
  - Our previous solution

The problem as a PSP

- Reordering can be seen as finding a permutation of columns in a bitmap.
- Considering unary coding, addressing the minimization of each posting list separately, and omitting the first offset, the problem consists in finding the order which minimizes the average d-gap sum.

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

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

- This last form 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.
- These problems are NP-complete.

TSP and optimality

- The TSP is only a strategy for addressing the real problem

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

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.

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

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.

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

- Although TSP is a good strategy, it is necessary to experiment with heuristics based on real-cost functions.

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.