Practical Representations for Web and Social Graphs*

Francisco Claude[1] and Susana Ladra[2]

[1] University of Waterloo, Canada.
fclaude@cs.uwaterloo.ca
[2] Universidade da Coruña, Spain.
sladra@udc.es

1 Summary

Graphs are a natural way of modeling connections among Web pages in a network or people according to a criteria like friendship, co-authorship, etc. Many algorithms that compute and infer interesting facts out of these networks work directly over these graphs. Some examples of this are Connected components, HITS, PageRank, spam-detection, among others. In social networks, graph mining also enables the study of populations' behavior. Successful graph mining not only enables segmentation of users but also prediction of behavior. Link analysis and graph mining remains an area of high interest and development.

These human-generated graphs are growing at an amazing pace, and their representation in main memory, secondary memory, and distributed systems are getting more and more attention. Furthermore, space-efficient representations for these graphs have succeeded at exploiting regularities that are particular to the domain. In the case of Web graphs the main properties exploited are the locality of reference, skewed in/out-degree, and similarity of adjacency lists among nodes of the same domain. In social networks, there is a tendency towards forming cliques in the network.

Hence, representing, compressing and indexing graphs become crucial aspects for the performance and in-memory processing when mining information from those graphs. In this work we focus on obtaining space-efficient in-memory representations for both, Web graphs and social networks.

We first show how by just partitioning the graph and combining two existing techniques for Web graph compression, k^2-trees [Brisaboa, Ladra and Navarro, SPIRE 2009] and RePair-Graph [Claude and Navarro, TWEB 2010], exploiting the fact that most links are intra-domain, we obtain the best time/space trade-off for direct and reverse navigation when compared to the state of the art. Our proposal, which is called k^2-partitioned, splits the graph in $t + 1$ pieces, the first t ones correspond to sub-graphs formed by groups of domains. The last piece contains all the edges that point from one of the t sub-graphs to another one. This combination allows us to obtain the best cases for the k^2-tree, since most of the compression is gained inside domains and the query time is good when the matrix is dense. For the representation of the internal links we

* This work has been funded by NSERC, David R. Cheriton Scholarships program, Ministerio de Ciencia e Innovación (grants TIN2009-14560-C03-02 and CDTI CEN-20091048) and Xunta de Galicia (grant 2010/17).

A. Ruíz, L. Iribarne (Eds.): JISBD'2012, pp. 81-82, ISBN:978-84-15487-28-9.
Jornadas SISTEDES'2012, Almería 17-19 sept. 2012, Universidad de Almería.

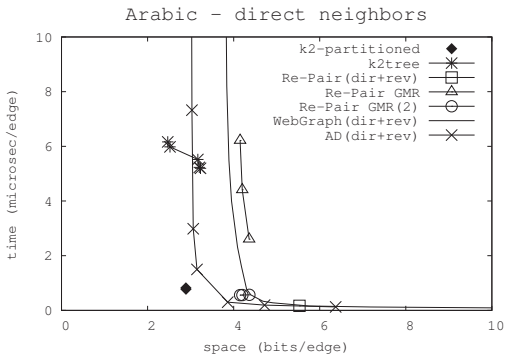

Fig. 1. Space/time trade-off to retrieve direct neighbors for several Web graphs.

use the k^2-tree and we represent external links using the approximate version of the Re-Pair compression method.

The experimental results show that for Web graphs we obtain a structure that improves upon the state of the art, achieving the best trade-off when we require both, direct and reverse navigation. This can bee seen in Figure 1, which shows the space/time trade-off for retrieving direct neighbors the large Web graph `arabic-2005` from the *WebGraph* project (reverse neighbors behave similarly). We measure the average time efficiency in μs/edge. We compare our compact representation, k^2-partitioned, with the original k^2-tree and other representations of the state of the art, including Re-Pair, Boldi and Vigna, and Apostolico and Drovandi's methods.

We can observe that our new representation k^2-partitioned competes successfully with the other compression methods of the literature, especially for larger graphs. It achieves very compact spaces, smaller than the rest of the techniques except for the k^2-tree, which can obtain slightly better spaces at the expense of degrading its time efficiency in orders of magnitude. Hence, our proposed technique achieves the best space/time trade-off for Web graph representation when direct and reverse navigation are required. Its simplicity contrasts with the remarkable results obtained.

We also study alternatives to compress social networks, where splitting the graph to achieve a good decomposition is not easy. For this case, we explore a new proposal for indexing MP_K linearizations [Maserrat and Pei, KDD 2010], which have proven to be an effective way of representing social networks in little space by exploiting common dense subgraphs. In the domain of social networks, our proposal improves upon previous results, both in theory and practice, showing that it constitutes a competitive index for social networks. Our implementations are available at `http://webgraphs.recoded.cl/`.

References

1. Claude F., Ladra S. Practical Representations for Web and Social Graphs. Proceedings of the 20th ACM International Conference on Information and Knowledge Management, CIKM'11. Glasgow, Scotland, UK. ACM Press, pages 1185–1190 (2011).

Reducción de la Complejidad Externa en Búsquedas por Similitud usando Técnicas de Clustering

Luis G. Ares, Nieves R. Brisaboa, Alberto Ordoñez, Oscar Pedreira*

Laboratorio de Bases de Datos, Universidade da Coruña
Campus de Elviña s/n, 15071, A Coruña, Spain
{lgares,brisaboa,alberto.ordonez,opedreira}@udc.es

Resumen La búsqueda por similitud tiene como finalidad determinar los objetos más semejantes o cercanos a uno dado. Los espacios métricos constituyen un modelo matemático que permite formalizar dicha búsqueda y que han dado lugar a diversos métodos, que tienen como objetivo principal reducir el número de evaluaciones de la función de distancia y el tamaño del índice. Las soluciones existentes son métodos basados en pivotes, que obtienen un número reducido de evaluaciones pero requieren cantidades importantes de espacio, y métodos basados en clustering, que necesitan poco espacio pero incrementan el número de evaluaciones. En este trabajo presentamos una nueva estrategia de clustering con sus algoritmos para búsquedas por rango y kNN que, reduciendo progresivamente el tamaño del cluster, disminuye significativamente la complejidad externa, un componente de la complejidad de los métodos existentes, con lo que se reduce el número de evaluaciones de la función de distancia.

Palabras Clave: Búsqueda por similitud, Espacios métricos, Reducción de cluster.

1 Introducción

El tratamiento de la información durante las últimas décadas ha abarcado toda clase de actividades humanas, originando tipos de datos complejos y un volumen de información en constante crecimiento. Entornos como los sistemas multimedia, la biología molecular y los sistemas de seguimiento y de recomendación, como los utilizados en las actividades industriales, financieras, sanitarias y sociales, ofrecen múltiples ejemplos de datos denominados *semiestructurados* y *no estructurados* donde los criterios de búsqueda no se basan en la exactitud, como ocurre con los tipos de datos tradicionales, y sí

* Trabajo parcialmente financiado por: Ministerio de Ciencia e Innovación (PGE y FEDER) refs. TIN2009-14560-C03-02, TIN2010-21246-C02-01, y ref. AP2010-6038 (programa FPU) para Alberto Ordóñez Pereira, y por Xunta de Galicia refs. 2010/17 (Fondos FEDER), y 10SIN028E.

A. Ruíz, L. Iribarne (Eds.): JISBD'2012, pp. 83-96, ISBN:978-84-15487-28-9.
Jornadas SISTEDES'2012, Almería 17-19 sept. 2012, Universidad de Almería.