Generalize geographic information to combine IR results

Damien Palacio, Christian Sallaberry, and Mauro Gaio

LIUPPA, Université de Pau et des Pays de l'Adour,
Avenue de l'Université, BP 1155, F-64013 Pau cedex
{damien.palacio, christian.sallaberry, mauro.gaio}@univ-pau.fr

ABSTRACT

Our contribution is dedicated to geographic information contained in unstructured textual documents. In order to combine spatial, temporal and topical IR results, we propose here a generic approach to homogenize information representations and relevance calculation formulae.

INTRODUCTION

We consider the generally accepted hypothesis that Geographical Information (GI) is made up of three components namely spatial, temporal and topical [1]. A typical GI sample contained in unstructured textual document is: "Fortified towns in South London suburbs during the 13th century." To process this textual unit, we claim that each of its three components should be treated independently, as is put forth by [2].

IR oriented research works define generalization as a stemming process of words in order to gather and weight them [3]. One of the most popular models developed in textual-based IR research is the vector space model [4] within which the content of each document can be approximately described by a vector of (content-bearing). This model is well accepted as an effective approach in modeling topical subspace.

Thereby, we propose to extend such a term-based matrix to a tile-based matrix in order to generalize topical, spatial and temporal indexes. This consists in rearranging geographic information into a uniform representation form, the tile, and to compute each tile occurrence frequency in the documents. Then, in a retrieval process, we compute statistics generally used for full-text IR to calculate dimension relevancy scores for each resulting document.

TILE-BASED GENERALIZATION APPROACH

We propose to adapt IR methods like representation generalization (lemmatization, truncation) and statistics approaches to heterogeneous information like spatial or temporal. So we propose a generic generalization approach that we will use for geographic information.

Tiling process consists in a "rasterization" of the information. But as the representation obtained by a segmentation are not necessarily represented by a regular grid, we chose to call them "tiles". Our approach consists on segmenting the space including all initial representations in a set of subspaces, the tiles. So, for one-dimension representations, tiling is one-dimension; for two-dimensions representations, tiling is two-dimensions, and so on until n dimensions. Thus this approach uses an existing index and generates a new index containing tiles.

Now we will describe formally this approach. To the domain[1] O included in space R^n corresponds a domain T included in space R^n[2]. The domain O is constituted of a set of objects $O_1, ..., O_p$ and the domain T is constituted of the union of \$$m$\$ subspaces (the tiles). For each subspace of T that intersect one or more objects of O, we keep the number of intersections (N_{Ti}).

[1] A domain is a finite set of values.

[2] This superset models spaces of dimension 1, 2 or more.

$$\begin{aligned} O \subseteq \mathbb{R}^n &\longrightarrow T \subseteq \mathbb{R}^n \\ O &= \{O_1, O_2, O_3, \ldots, O_p\} \\ T &= \bigcup_{i=1}^{m} T_i \\ N_{T_i} = \mid T_i \mid T_i \cap O_j \neq \varnothing \quad \forall j &= 1, \ldots, p \mid \\ \text{with } \mid x \mid &\text{ the cardinality of } x \end{aligned} \tag{1}$$

After the tiling chosen, we can weight tiles using frequency-based approaches. To calculate a tile frequency, we propose two discrete approaches. First, binary frequency consists on counting the number of initial representation of the information (object) that intersects it, while keeping in mind that one object can intersect several tiles. Second, proportional frequency consists on: according to the ratio overlay object/tile, a tile frequency is incremented by a value between 0 and 1. Table 1 details the two types of frequencies.

Binary frequency	$freq(T_i) = \sum_{j=1}^{p} freq(T_i, O_j)$
Proportional frequency	$freqP(T_i) = \sum_{j=1}^{p} freq(T_i, O_j) * \frac{Ar(T_i,O_j)}{Ar(T_i)} * \frac{1}{NbTiles(O_j)}$

Table 1*: Frequency formulae (freq($T_{i,}O_j$): object $O_{\_j}$ frequency in tile, T_i (number of intersection), Ar(T_i,O_j): object $O_{j\$}$ area on tile T_i, Ar(Ti): tile T_i area, NbTiles(O_j): number of tiles intersected by object O_j)*

This general tile-based indexing strategy allows the implementation of state-of-the-art models for relevance score computation based on tiles occurrence frequencies within documents like TF.IDF or OkapiBM25.

EXPERIMENTS

We used MIDR_2010 corpus (travelogues digitalized) [5] (the same we used for spatial queries) and we submitted 35 temporal queries to our classical temporal IRS (baseline) and our tile-based temporal IRS. Tile-based temporal IRS also tried different regular and calendar (explicit) segmentations. In the same way, it tested these segmentations with different weighting formulae. Table 2 shows that tile-based temporal IR gives results that are quite similar to those retrieved by the baseline: i.e. the month segmentation associated to the TFp weighting formula is well-suited to our cultural heritage corpus.

GIRS	MAP	Improvement
Classical Temporal IRS	**0.93**	0,0 %
Tile-based Temporal IRS	0.92	−0,5 %

Table 2*: Tile-based temporal IR effectiveness (Mean Average Precision)*

Here, the loss of accuracy due to the gathering process is not fully compensated: the temporal information distribution within our corpus (no more than one temporal information in a paragraph) explains why the temporal information frequency computation cannot improve the baseline results.

CONCLUSION

This paper proposes a generalization approach that we apply to geographic information. Furthermore, experiments showed the effectiveness of our solution: a proportional spatial or temporal tiles frequency processing associated with a proportional document units relevancy computation gives equivalent (for temporal) or better results (for spatial) for monodimensional tile-based IR modules than for monodimensional standard IR baselines. We have proposed this generalization approach in order to combine such geographic IR results. It has allowed us to test operator to combine spatial, temporal and topical (terms) and we showed that the combination really improve IR results (+66,3%) [6]. We plan to apply this generalization approach for topical dimension: projecting term on concepts from ontologies.

REFERENCES

1. Longley, P.A., Goodchild, M.F., Maguire, D.J., Rhind, D.W.: Geographic Information Systems and Science. John Wiley & Sons (2005)

2. Clough, P., Joho, H., Purves, R.: Judging the Spatial Relevance of Documents for GIR. In: ECIR'06: Proceedings of the 28th European Conference on IR Research. Volume 3936 of Lecture Notes in Computer Science., Springer (2006) 548-552

3. Spärck Jones, K.: A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation 28(1) (1972) 11{21

4. Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval. McGraw-Hill (1983)

5. Palacio, D., Sallaberry, C., Gaio, M.: Normalizing Spatial Information to Improve Geographical Information Indexing and Retrieval in Digital Libraries. In: ISGIS'10: Proceedings of the Joint International Conference on Theory, Data Handling and Modelling in GeoSpatial Information Science proceedings. (2010) 229-234

6. Palacio, D., Cabanac, G., Sallaberry, C., Hubert, G.: Measuring Effectiveness of Geographic IR Systems in Digital Libraries: Evaluation Framework and Case Study. In Lalmas, M., Jose, J., Rauber, A., Sebastiani, F., Frommholz, I., eds.: ECDL'10: Proceedings of the 14th European Conference on Research and Advanced Technology for Digital Libraries. Volume 6273 of LNCS., Springer (September 2010) 340-351