A toponym-based dual vector for topical relevance calculation in focused spatial crawling

Dongyang Hou[1,2], Hao Wu [1], Jun Chen [1]

[1] National Geomatics Center of China, Beijing 100830, China

[2] School of Environment Science and Spatial Informatics , China University of Mining and Technology, Xuzhou,Jiangsu,221116, China

1. Introduction

Focused crawler is a Web crawler that tries to download only pages that are relevant to a given topic of interest (Siemiński 2009, Almpanidis 2011). That is to say, it is necessary for focused crawler to calculate relevance between pages and specific topic (Rungsawang, 2005). Recently, the specific topic involving spatial information especially toponyms such as the topic about the Diaoyu Islands conflict between China and Japan has become much more, because there are about 18 percent webpages describing localization information and 70 percent webpages containing geographic information webpages in the worldwide webpages (Zhou et al. 2005, Hill 2009). For the given topic involving toponyms, it means that focused crawler should ensure the downloaded pages are not only relevant with common topic and also relevant with the toponyms, which requires more accurate topical relevance calculation.

At present, most researchers adopt Vector Space Model (VSM) (Fox 1984, Batsakis et al. 2009) or improved VSM (Yang et al. 2010, Li 2008) to calculate the topical relevance. In these methods the given topic and document are regarded as one keywords vector, which means that toponyms are consider as common topic. Considering toponyms as common topic may reduce the crawling accuracy, because a toponym can involve different common topics and a common topic also can relate to many toponyms.

To solve the problem, we make a difference between toponyms and common keywords according to the idea of Geographic Information Retrieval (Purves 2007) and develop a toponym-based dual vector for topical relevance calculation.

2. The toponym-based dual vector

In traditional VSM, a topic T and a document D are separately represented as a single vector V_T and V_D, as shown in equation (1) and (2), in which toponyms are regarded as common keywords.

$$V_T = (w_{t1}, w_{t2}, \cdots, w_{tN}) \tag{1}$$

$$V_D = (w_{d1}, w_{d2}, \cdots, w_{dN}) \tag{2}$$

Where $\mathrm{w_{tn}}$ and $\mathrm{w_{dn}}$ represent the weight of Nth keyword in topic T and document D respectively.

In the paper, toponyms are regarded as an independent vector and the rest common keywords are considered as another vector. We call the method toponym-based dual vector.

In toponym-based dual vector, a topic T is represented as toponyms vector TV_T and common keywords vector CKV_T, as shown in equation (3) and (4). Correspondingly, a document D can be expressed as TV_D and CKV_D, as shown in equation (5) and (6).

$$\mathrm{TV_T = (w_{t1}, w_{t2}, \cdots, w_{tm})} \tag{3}$$

$$\mathrm{CKV_T = (w_{t1}, w_{t2}, \cdots, w_{tn})} \tag{4}$$

$$\mathrm{TV_D = (w_{d1}, w_{d2}, \cdots, w_{dm})} \tag{5}$$

$$\mathrm{CKV_D = (w_{d1}, w_{d2}, \cdots, w_{dn})} \tag{6}$$

Where $\mathrm{w_{tm}}$ and $\mathrm{w_{dm}}$ represents the weight of mth toponym in topic T and document D respectively. $\mathrm{w_{tn}}$ and $\mathrm{w_{dn}}$ indicts the weight of nth common keyword in topic T and document D separately.

The weight $\mathrm{w_{tm}}$ and $\mathrm{w_{tn}}$ can be set by manual or be calculated in corpus in the paper. The weight $\mathrm{w_{dm}}$ and $\mathrm{w_{dn}}$ are often calculated through term-frequency (tf) algorithm or term frequency-inverse frequency (tf-idf) algorithm. Because computing inverse document frequency (idf) weights during crawling may be problematic (Batsakis 2009), the tf algorithm is employed in the paper. Therefore, $\mathrm{w_{dm} = tf_{dm}}$ and $\mathrm{w_{dn} = tf_{dn}}$ where $\mathrm{tf_{dm}}$ and $\mathrm{tf_{dn}}$represent the occurrence of mth toponym and nth common keyword in the document D.

3. Relevance calculation utilizing toponym-based dual vector

In the paper, the topic T and document D are both represented by two vectors, therefore, the relevance calculation is carried out through two-steps, as shown in figure 1.

The first step is calculating the relevance between common keywords vector of topic and document. For common keywords vectors $\mathrm{CKV_T}$ and $\mathrm{CKV_D}$, the relevance between them is calculated by cosine similarity function (Hao et al. 2011) as equation (7) shows.

$$\mathrm{Sim(CKV_T, CKV_D)} = \frac{\sum_{i=1}^{n} w_{ti} * w_{di}}{\sqrt{\sum_{i=1}^{n} w_{ti}^2 * \sum_{i=1}^{n} w_{di}^2}} \tag{7}$$

If $\mathrm{Sim(CKV_T, CKV_D)}$ is greater than the specific threshold σ_1, the next step will be implemented, if otherwise, the document will be abandoned.

The second step is computing the relevance between theirs toponyms vector. Similarly, the relevance between $\mathrm{TV_T}$ and $\mathrm{TV_D}$ is calculated as shown in equation (8).

$$\mathrm{Sim(TV_T, TV_D)} = \frac{\sum_{i=1}^{m} w_{ti} * w_{di}}{\sqrt{\sum_{i=1}^{m} w_{ti}^2 * \sum_{i=1}^{m} w_{di}^2}} \tag{8}$$

If $\mathrm{Sim}(TV_T, TV_D)$ is greater than the specific threshold σ_2, the document is judged to be relevant with the given topic.

At last, the finial relevance $\mathrm{Sim}(V_T, V_D)$, which will be utilized in predicting URL queue priority, is the weighted average of $\mathrm{Sim}(CKV_T, CKV_D)$ and $\mathrm{Sim}(TV_T, TV_D)$:

$$\mathrm{Sim}(V_T, V_D) = a * \mathrm{Sim}(CKV_T, CKV_D) + b * \mathrm{Sim}(TV_T, TV_D)$$

Where a and b are weighted factors, and they satisfy $a + b = 1$. In the paper, a is assigned to 0.4 and b is assigned to 0.6.

Note that if the specific topic does not contain toponym, the method will be the same to the traditional method, that is only the first step is implemented.

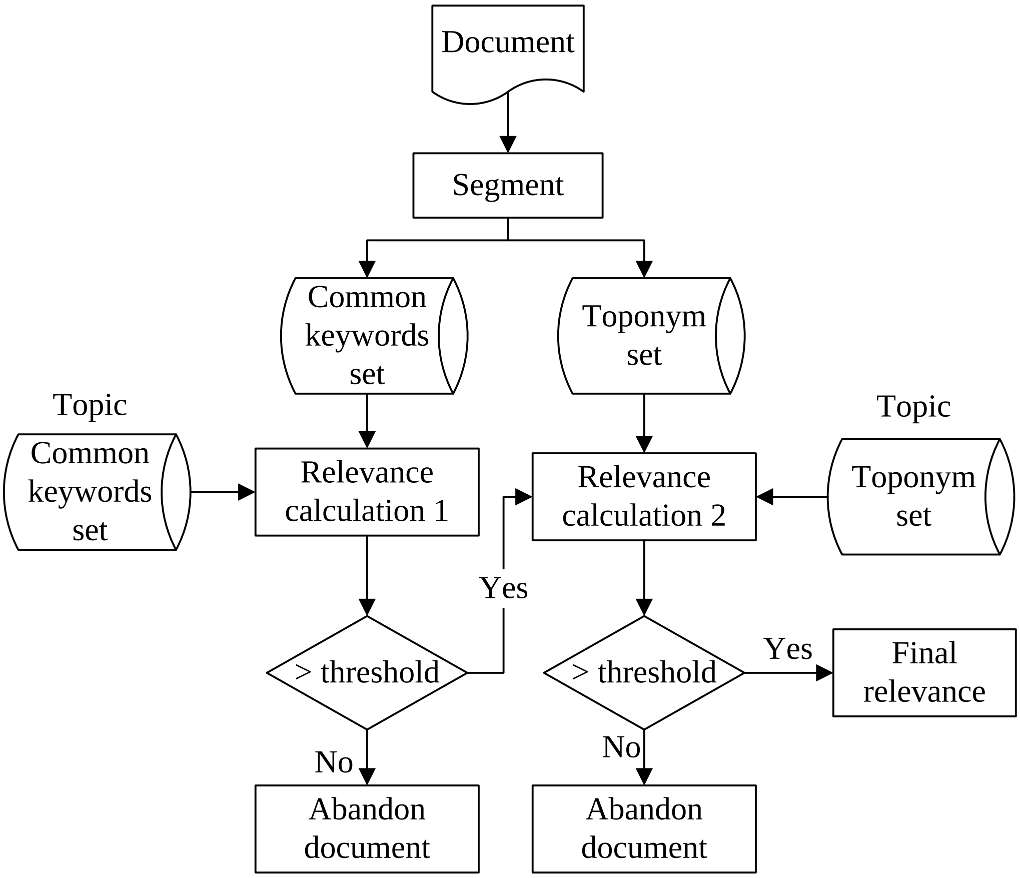

Figure 1. Frame diagram of relevance calculation utilizing toponym-based dual vector

4. Experiment

In order to validate the toponym-based dual vector for topical relevance calculation, a simple experiment is implemented. In the experiment, the given topic is the Diaoyu Islands conflict between China and Japan. Five common keywords CK_T and five toponyms T_T which are extracted by some sample Chinese news represent the specific topic.

$$CK_T = \left\{ \begin{matrix} (\text{Cruise}, 0.06), (\text{On islands}, 0.06), (\text{Conflict}, 0.05), \\ (\text{Protest}, 0.06), (\text{Government surveillance vessel}, 0.25) \end{matrix} \right\}$$

$$T_T = \left\{ \begin{matrix} (\text{Diaoyu Islands}, 1), (\text{Japan}, 0.8), \\ (\text{China}, 0.6), (\text{Taiwan}, 0.1), (\text{Hongkong}, 0.1) \end{matrix} \right\}$$

In table 1, four test documents D1, D2, D3 and D4 are represented by the ten keywords, where D1 and D3 are relevant with the given topic, D2 is about the Huangyan Island conflict between China and Philippines, D4 is about the economic comparison between China and Japan, and GSV is the abbreviation of Government surveillance vessel.

In table 2, the relevance between the four documents and topic is calculated through traditional method and the proposed method.

	Cruise	On island	Conflict	Protest	Gsv	Diaoyu Islands	Japan	China	Taiwan	Hong kong
D1	3	0	0	0	3	1	3	2	0	0
D2	7	0	5	0	7	0	0	8	0	0
D3	0	2	4	2	0	37	23	10	5	1
D4	0	0	1	0	0	0	37	31	0	0

Table 1. Document representation for four test documents

	$\mathrm{Sim}(\mathrm{T}, \mathrm{V})$	$\mathrm{Sim}(\mathrm{CKV}_\mathrm{T}, \mathrm{CKV}_\mathrm{D})$	$\mathrm{Sim}(\mathrm{TV}_\mathrm{T}, \mathrm{TV}_\mathrm{D})$
D1	0.62	0.71	0.87
D2	0.31	0.79	0.42
D3	0.97	0.64	0.97
D4	0.69	0.36	0.70

Table 2. Results of the relevance calculation utilising two methods

As shown in table 2, if all threshold values are 0.5, only D2 can be filtered out in traditional method, but in the new strategy D4 can be filtered out in the first step and D2 can be filtered out in the second step. Therefore, the new method can improve the accuracy of relevance judgments.

5. Conclusions

For the given topic involving toponyms, we develop a toponym-based dual vector for topical relevance calculation. In the method, the toponyms are regarded as an independent vector and the relevance calculation is carried out through two steps. Through a simple experiment, we can conclude that the proposed method can improve the accuracy of relevance judgments.

In the future, we will carry out some complex experiments to further validate the method’s performance.

6. References

Almpanidis G, Kotropoulos C and Pitas I, 2007, Combining text and link analysis for focused crawler-An application for vertical search engines. *Information Systems*, 32:886-908.

Batsakis S, Petrakis E G M and Milios E, 2009, Improving the performance of focused web crawlers. *Data & Knowledge Engineering*, 68:1001-1013.

Fox E A, 1984, Extending the boolean and vector space models of Information Retrieval with p-norm queries and multiple concept types, *Dissertation Abstracts International Part B: Science and*

Engineering, NYC: Cornell University, 44(9):386.

Hao H W, Mu C X, Yin X C and Li S et al., 2011, An Improved Topic Relevance Algorithm for Focused Crawling. *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Anchorage, AK, 850-855.

Hill L L, 2009, Georeferencing: The Geographic Associations of Information. The MIT Press.

Lu C S, 2011, Thematic VSM based on ontology semantic tree. *Computer Systems & Applications*, 20(10):44-48. (in Chinese)

Purves R S, Clough P, Jones C B and Arampatzis A et al., 2007, The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the Internet. *International Journal of Geographical Information Science*, 21(7): 717-745.

Rungsawang A, Angkawattanawit N, 2005, Learnable topic-specific web crawler, *Journal of Network and Computer Applications*, 28(2):97-114.

Siemiński A, 2009, Using WordNet to Measure the Similarity of Link Texts, In: Nguyen N T, Kowalczyk R and Chen S M (eds), *Proceedings of First International Conference, ICCCI 2009*, Wrocław, Poland, October, 720-731.

Yang X, Sui A N and Tang Z K, 2010, Topical Crawler Based On Multi-Lev el Vector Space Model and Optimized Hyperlink Chosen Strategy, In: Sun F, Wang Y, Lu J, Zhang B, Kinsner W and Zadeh L A (eds.) , *Proceedings of the 9th IEEE International Conference on Cognitive Informatics (ICCI 2010),* 430-435.

Zhou Y H, Xie X, Wang C and Gong Y C et al., 2005, Hybrid index structures for location-based web search. *Proceedings of the 14th ACM international conference on Information and knowledge management*. Bremen, Germany, 155–162.