Research on Geospatial Information Grab and model Evaluation based on Micro-blog

Xiaochun Liu

Information Engineering University, Zhengzhou, China

253807321@qq.com

Keywords: Geospatial information; Micro-blog; Acquisition System.

Abstract. Surfing on the internet, people can read abundant useful information worldwide. Meanwhile the readers have to receive the huge number of redundant online information either, especially in the field of geospatial information. This paper aims firstly to analyze the problems on the retrieval of geospatial based on micro-blog, with running multiple simulations of the models, the performance of geospatial Information Acquisition System is analyzed, and the proposition for improving geospatial Information Acquisition System is presented.

Introduction

The geospatial data including both attribute information and geometry information, which is special and unique from other kinds of data, can only be represented by the description of texts and geometric graphs. And so far, the main focus on the retrieval of geospatial information is the description and matching of texts, while less focus on that of geometry information. Micro-blog as a network platform of information sharing and communication has been widely used in recent years. Until 2011, active persons using micro-blog in China reached 19.5 billion.

Micro-blog network has attached the attention of some scholars not including the geography research. Micro-blog extraction characteristics make it a good platform of network information flow spatial distribution, the platform built by the real user relationship makes virtual cyberspace materialized, and each user is coincided with the geographical spatial phase coupling. Micro-blog is the most influential and most high-profile platform. In the first quarter of 2011, micro-blog users account for 57% of total domestic users. This paper established vector space model based on Sina micro-blog, realizing the real-time geographic spatial information acquisition.

Web Crawler (Web Crawler), also known as the (Web spiders) or Web information collector, is an automatic extraction procedure of Web applications, is an important part of the search engine. Traditional web crawler begin from one or more initial web URL, to obtain the initial web page link URL, in the process of crawl the web, from the current web page for new URL into the queue, until meeting certain conditions precedent.

Spatial information reflects the characteristics of geographical entity spatial distribution of information. The spatial distribution characteristics, including spatial data, the entity's location, shape and the spatial relationship between entities, K domain space structure, etc. where shape, space and regional spatial structure information is often seen as graph, table, file etc, and it is difficult to unity the data format and recognition, so it is difficult to build index, conventional search.

Study on the vector space model of retrieval

Vector space model (VSM: Vector Space Model) proposed by Salton and others in twentieth Century 70 years, it is the basic idea of each text and query contains some features independent properties reveal its content, and each feature attributes can be regarded as a dimension vector space, then the text can be expressed as a collection of these attributes, ignoring the complex relationship between paragraphs, sentences and words in the text structure. At the same time, given the feature weight vocabulary certain (weight), anti should vocabulary in the importance and the value of the contents of the file identification, this value is called the indexing vocabulary "significant value (Term Significances)" or "weight", by the lexical statistics calculate the document and to, such as: the

feature words appear frequency (Term frequency, TF). Vector of each file is in fact all the document feature through a combination of computing, called "the document feature item vocabulary matrix". And then all of the document vector based on specific computing methods of similarity measure between each other.

Vector space retrieval model can be described as $I = (D, T, Q, F, R)$ Among them: $D = \{d_1, d_2, ...d_n\}$ As a collection of text, n text collection number; $T = \{t_1, t_2, ...t_n\}$ Set as a feature, m feature of all. A text m feature indexing can be represented as a vector space $d_i = \{w_{i1}, w_{i2}, ...w_{im}\}, i = 1, 2, ..., n$, w_{ij} is characteristic t_j for the text d_i of the weight, if the weight value w_{ij} is 0, indicating t_j that it is not appeared in d_i, $Q = \{q_1, q_2, ...q_m\}$ for the query set, a query q_r can be represented by vectors $q_r = \{q_{r1}, q_{r2}, ...q_{rm}\}$, q_{rj} is a characteristic to t_j the query q_r weights, if the weight value q_r is 0, indicating that t_j is not appeared in q_r.

Further definition: Frequency tf_{ij}: t_j is the feature for text d_i appear in the frequency;

Inverse document frequency word idf_i (inverse document frequency): the word in the quantitative distribution of document collection, the calculation $\log(N / n_k + 0.5)$ is usually, where N is the total number of document centralized, n represents a number of documents containing K, called the document frequency of the term.

The normalization factor: in order to reduce the inhibitory effect of high frequency characteristics of individual word on other low-frequency feature words, the standardization of components.

Based on the above three factors to term weighting formula(1):

$$w_{ik} = \frac{tf_{ik} \log(N / n_k + 0.5)}{\sqrt{\sum_{k=1}^{n} (tf_{ik})^2 \times [\log(N / n_k + 0.5)]^2}} \tag{1}$$

The similarity between the text and the query can be used to measure the distance between two vectors. There are many kinds of calculating method of similarity, commonly used methods of inner product, Dice coefficient, Jaccard coefficient and cosine coefficient, usually uses the cosine coefficient method, namely the cosine of the angle between two vectors to represent the similarity between the text and the query $Sim(d_i, q_j)$, see equation (2). Cosine similarity calculation method is a normalization, the angle between the two vectors of the smaller, the greater the degree of correlation between documents, correspondence $\cos$ is higher. Two vector included angle cosine is equivalent to their standard vector inner product unit length, it reflects the similarity term component two vector of relative distribution.

$$Sim(d_i, q_j) = \cos\theta = \frac{\sum_{k=1}^{n} w_{ik} \times w_{jk}}{\sqrt{(\sum_{k=1}^{n} {w_{ik}}^2) \times (\sum_{k=1}^{n} {w_{jk}}^2)}} \tag{2}$$

Experimental results and analysis

Data. The experimental data are collected through a self- extracting and Sina micro-blog open API. The data set includes 180,000 posts information. After cleaning up the data set after removal of junk information used in this experiment. First it should clean out micro-blog of fewer than 10 characters, then get rid of less than three nouns micro-blog , and finally remove to begin with @ micro-blog information, because these micro-blog user's private conversations are actually unsuitable recommendations to other users.

Algorithm evaluation and methods comparison

This section through experimentally compared int λ and based on cosine similarity and label vector algorithm performance and evaluation mechanism using a variety of automatic evaluation of the int λ effect. Experiment randomly select 200 users from a centralized data , for each user, select its 90 % of the published micro-blog as a measure of the user's own interest data , and the remaining 10% of the part and the other all users of the micro-blog mixed together to form a paper test data sets. Assuming paper recommendation system will test, if the data set 10% of the user's own micro-blog recommended to the user is called the right recommendation.

This paper uses a standard evaluation methods for geospatial information retrieval as a mechanism for the evaluation of information retrieval models. We calculate for each user:

P@k, precision arithmetic refers to the proportion of pre- k micro-blog of right recommended;

S@k, the success rate is at least in the first k posts , there is a correct proportion of micro-blog;

MRR, average ranking algorithm is to last in the final sorted sequence results in reciprocal Q, and each of the article the correct location of the micro-blog the average (if (3) shown above), high MRR indicating a higher accuracy of the algorithm.

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i} \tag{3}$$

Firstly, in order to calculate λ value optimization in int λ, this experiment from the data set 200 users using the above parameters are chosen for testing. Figure 1 P@k, under different values of λ is displayed the performance results int_λ .

Table 1 Comparison of methods of int0.9, cosine similarity and label

	P@1	P@3	P@5	MRR
Int0.9	0.69	0.60	0.53	0.73
cosin	0.30	0.26	0.22	0.46
Tag	0.33	0.28	0.21	0.51

Results show that when λ values is close to 1, p score has more weight, and corresponding to the model in the evaluation of algorithms is also better. In fact, the maximum value P @ k and S @ k is obtained when the λ = 1. when λ =0.9 in the values of λ, MRR shows a downward trend, which corresponds to the correct micro-blog's ranking is on the rise. Therefore, when λ = 0.9 , int_λ on the accuracy and success rates are doing well and will be interesting micro-blog arranged in a higher position .

Conclusion

This paper constructed space vector model of information acquisition system based on micro-blogging, can understand and aware of the operation process of the system and the relationship between each component of image, but test model needs to be further completed, spatial information service oriented to parse and found services, its application is far from ideal state, it is still very difficult to the service retrieval and check. Due to spatial data application is very broad, involves the extensive information content, at present we only can obtain for specific application data, unable to establish a unified model to realize all of the data processing.

Reference

[1] Stefan Hinz, Albert Baumgartner. Automatic Extraction of Urban Road Networks from Multi-View Aerial Imagery [J]. ISPRS Journal of Photo and Remote Sensing. 2003(58):83-98.

[2] W. Li, C.Yang, etc. Semantic-Based Web Service Discovery and Chaining For Building an Active Spatial Data Infrastructure [J]. Computer & GeoSciences. 2011(37):1752-1762.

[3] Ming Fu, Data mining research based on the space of Web [D]. South university. 2004.

[4] Bin He, Accessing the Deep Web [J].Communications of the ACM. 2007(50):94-101.