On-line Event Detection from Web News Stream [Good for Related Entity Identification]

4.2. Story Model with News Elements

News stories are reports for specified events with certain time, places, persons or organizations. From this point of view, time, place and person elements are very important and distinguished for the content of a piece of news. Once we confirm these elements, we can limit the event to a very small range. Furthermore, these important elements are named entities and could be extracted with named entity recognition.

We propose a new story model SNE(Story with Named Entity) to describe news stories based on previous analysis. Then, a piece of news story could be represented as s=(D,Pl,Pe,C). D, Pl and Pe represents the date, place or location, and the involved person or organization(if there is no specific person appearing in the news story, the organization name is always taken as the subject) respectively. C is the original text content of the story. To extract such news elements from stories, we introduce named entity recognition. Since date entities always appear regularly, we extract them based on established rules. And this extraction process is similar to the string matching with regular expressions. ICTCLAS(Institute of Computing Technology Chinese Lexical Analysis System, www.nlp.org.cn) is adopted to accomplish recognition of persons, places and organizations.

Usually, there are always more than one date appearing in one piece of news, so does place, person and organization. In addition, even the same date or place would appear more than once in a story sometimes. Hence, we usually get numerous news elements after extraction. However, these elements are not as important as each other. In general, elements which make the story different even distinguished should play more important roles in the detection task. Then we try to figure out the importance of each extracted news element, based on its ability of discriminating one story from the others. We take two factors into consideration, one is frequency of elements, another is the first appearing position. Compared with the same kind of entities, an news element should be assigned with bigger weight when it has higher frequency or is mentioned at the very beginning of the story.

Suppose that *d* is a date in story *s*, then the weight of *d* is defined as

$$w(d,s) = \log[(1+\frac{C_d}{N}) \times \frac{1+(N-L_d)}{\sum_{j=1}^{N} j}]$$

C_d means times that *d* appearing in *s*, N is the total count of date entities in *s*, so Cd/N represents the frequency for *d* in *s*. We organize all date entities in *s* according to their appearing order and construct an ordered sequence. Then L_d means the first position number for *d* in the sequence.

Computation and processing of place, person and organization entities are the same as date entities. We obtain ranked results of weighted entities, then choose the element with the highest score in each entity type to form SNE model.