A Fine-Grained Digestion of News Webpages through Event Snippet Extraction

Rui Yan
Dept. of Computer Science
Peking University, China
r.yan@pku.edu.cn

Liang Kong
Dept. of Machine Intelligence
Peking University, China
kongliang@pku.edu.cn

Yu Li
School of Computer Science
Beihang University, China
carp84@gmail.com

Yan Zhang *
Dept. of Machine Intelligence
Peking University, China
zhy@cis.pku.edu.cn

Xiaoming Li
Dept. of Computer Science
Peking University, China
lxm@pku.edu.cn

ABSTRACT

We describe a framework to digest news webpages in finer granularity: to extract event snippets from contexts. "Events" are atomic text snippets and a news article is constituted by more than one event snippet. Event Snippet Extraction (ESE) aims to mine these snippets out. The problem is important because its solutions may be applied to many information mining and retrieval tasks. The challenge is to exploit rich features to detect snippet boundaries, including various semantic, syntactic and visual features. We run experiments to present the effectiveness of our approaches.

Categories and Subject Descriptors

H.4.m [**Information Systems**]: Miscellaneous

General Terms

Algorithms, Experimentation, Measurement

Keywords

News digestion, event snippet extraction, web mining

1. INTRODUCTION

With a large volume of news webpages on the Web, news digestion increasingly becomes an essential component of web contents analysis. We investigate the problem of Event Snippet Extraction (ESE) to divide a news webpage into event-centered snippets. ESE is highly motivated because the event distilling improves retrieval experience by presenting only the relevant parts instead of the whole page and is of potential use in applications like discourse analysis. News clustering and classification can also be in accurate granularity with less jeopardized noises and so be content extraction and webpage deduplication. To sum up, fine-grained digestions by ESE open doors to wide use on Web.

ESE is related to traditional text segmentation which often fails to be event-oriented [1]. [2] proposes an introductive work but can still be polished: we consider more detailed elements such as rich semantic, syntactic and visual features.

*Corresponding author.

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

2. SNIPPET EXTRACTION

Based on the topic drift principle investigated in [2], we treat the sentence s with a timestamp as a potential head sentence (s_h) of an event snippet (S). Assume in the news document D ($D = \{s_1, s_2, \ldots, s_{|D|}\}$) each S can be represented as $<t{:}\{s\}>$, where t is the timestamp of s_h and $\{s\}$ is the set of sentences that belong to S. Suppose there are m snippets in D, $(\cup_{k=1}^{m} S_k) \subseteq D$ and $\forall i \neq j$, $S_i \cap S_j = \varnothing$. Original sentence sequence is preserved in snippets. Neighboring contexts tend to describe the same event due to the semantic consecutiveness of natural language discourse. A snippet expands by absorbing texts pertinent to the event.

2.1 Semantic Relevance

Intuitively, semantic relevance *Rel* of a pending sentence (s_p) to S can be measured by the probability being generated from the language model of the snippet (LM(S)), as defined in Equation (1). Sentences with low probability are clearly off-event and not related to the expanding snippet.

$$Rel(s_p, S) = p(s_p|LM(S)) = \left(\prod_{w \in s_p} \frac{\sum_{s_i \in S} tf(w, s_i) + \lambda}{(1+\lambda) \cdot \sum_{s_i \in S} |s_i|} \right)^{\frac{1}{|s_p|}} \quad (1)$$

λ is empirically set at 0.01 as a smoothing factor and $|s|$ is the size of sentence s. $tf(w, s)$ is the term frequency of word w in s. Equation (1) assumes all sentences are equally weighted while in fact some sentences have larger probability to be on-event than others in S. We denote such probability as sentence **significance**. Semantic, syntactic and visual features distinguish significance and we exploit them next.

2.2 Weighted Semantic Relevance

Distance Decay (DD). The tendency for contexts to agglomerate attenuates as distance becomes larger from head sentence s_h, i.e., a distance decay. According to our investigation and statistics in [2], the snippet length L follows a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$. Given $x = ||s_p||$ where $||s||$ is the offset of sentence s from s_h, distance decay $f_d(x)$ is:

$$f_d(x) = P(x < L) = \int_x^{+\infty} \mathcal{N}(t, \mu, \sigma^2) dt. \quad (2)$$

Temporal Proximity (TP). Re-mention of adjacent temporal information may strengthen event continuousness and raise significance, but huge time gap indicates separate events. Given $\Delta t = |t_n - t|$ where t_n is the new time contained in sentence s_t and t is from s_h, T_D as the time span of D, temporal proximity $f_t(x) = e^{-\alpha \times \frac{\Delta t}{T_D}} \times f_d(x - ||s_t||)$.

Named Entities (NE). Sentence with named entities (s_e) might indicate strong relevance if entities are connected by existing knowledge databases (e.g. WordNet or Wikipedia), but [2] assumed equal distance for all adjacent entities in hierarchical taxonomy structures. Leaf/lower level entities should be closer than general concepts from higher levels. Consider a fragment, <health [food safety, public health organization(Centers for Disease Control, World Health Organization)]>, (CDC, WHO) are closer than (food safety, public health organization). We model synonyms, hyponyms and hypernyms into entity distance. We assign a distance weight (w_e) to every entity $w_e = 1 + \sum_{e_k \in H(e)} w_{e_k}$ where $H(e)$ is the hyponym set of entity e. The distance from a hyernym e to one of its hyponym e_k is defined as:

$$dist = \sum_{e_k \in H(e)} \frac{w_{e_k}}{|H(e)|}. \quad (3)$$

The weight of leaf node is set as 1. *dist* and *weight* are measured separately and penalization costs more for category entities. Entity influence $f_e(x) = e^{-\beta \times dist} \times f_d(x - ||s_e||)$.

α, β are scaling factors. $f_d(x)$, $f_e(x)$, $f_t(x)$ affect sentence significance separately and there are more than one s_e or s_t in S. For snippet completeness we choose the maximum $f_e(x)$ and $f_t(x)$ and take the arithmetic average of the three.

Conjunctive Indicators (CI). Conjunctions such as "however", "so", etc. reflect the author's intention of a semantic bridge between the adjacent sentences, which raises sentence significance. For the sentence with these conjunctive indicators, we assume it shares the same significance with its neighboring sentence prior to it. The conjunctive influence is local and not accumulative to following texts.

$$sig(x) = sig(x-1) \quad \text{if } (s_x \cap s_{x-1}) \subseteq \text{CI}. \quad (4)$$

Layout Presentation (LP). The visual structure of the news article in the webpage can give some clues to the event atoms, since writing style implies event principles as well.

- *Line break.* When meet the tag of
 or <p>, the line break as the author's intention of topic drifting.
- *Visual Elements.* An inserted image, table or hyperlink (<img>, <a>, etc.) indicates similar effect as line breaks due to news writing style.

The effects of line break and visual elements are accumulative. After τ visual changes, the probability drops by $\prod_\tau (1 - r_i)$. r_i are not equal due to specific contexts but for simplicity we assume they are all r. Hence final $sig(.)$ is:

$$sig(x) = (f_d(x) + max\{f_e(x)\} + max\{f_t(x)\}) \times (1-r)^\tau / 3 \quad (5)$$

Combining Significance. Each sentence in snippet affects following sentences, either increasing or decreasing the significance. We apply $sig(.)$ in Equation (1) and obtain a weighted relevance score from all sentence pairs between s_p and sentences in the expanding snippet S. We add s_p into S when relevance exceeds a threshold.

$$p(s_p|LM(S)) = \left(\prod_{w \in s_p} \frac{\sum_{s_i \in S} sig(s_i) \cdot tf(w, s_i) + \lambda}{(1+\lambda) \cdot \sum_{s_i \in S} sig(s_i) \cdot |s_i|} \right)^{\frac{1}{|s_p|}} \quad (6)$$

3. EXPERIMENTS

In a 10-fold cross validation manner, we test our proposed approaches on a corpus of 1000 webpages from the *Xinhua News* website. There are on average 1.893 snippets per news document and for all snippets, $\mu = 6.97$, $\sigma = 2.11$. Golden standards are created by human annotators. α, β, r are set experimentally at 0.6, 0.5, 0.174 correspondingly. We stick to the *precision/recall* evaluation metrics in [2]. Figure 1 shows the experiment results of semantic relevance (SeRel) and weighted semantic relevance (WSeRel) compared with TextTiling proposed in [1], TTM and LGM proposed in [2]. The perfromance of different features is shown in Figure 2.

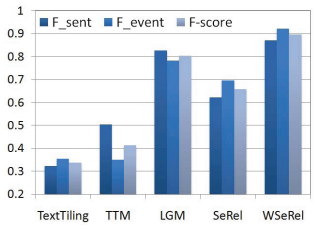

Figure 1: Performance

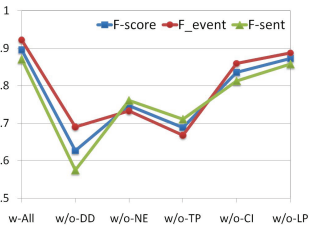

Figure 2: Features

WSeRel generally outperforms others. TextTiling shows significant weakness because it is not event-oriented. The contribution of **significance** is obvious (+26.56%) by comparing WSeRel with SeRel. DD is the most essential for snippet expansion. TP, NE, CI are also necessary. LP seems not to perform well due to misleading line breaks and visual noises. We present a system demonstration snapshot.

Figure 3: Fine-grained news digestion system demo.

4. CONCLUSIONS

We describe a fine-grained news digestion framework of ESE, utilizing semantic, syntactic and visual features. ESE is an on-going infrastructure work facilitating other researches. We show that our approach outperforms rival methods.

5. ACKNOWLEDGMENTS

This work is partially supported by NSFC with Grant No.60933004, 61050009 and 61073081.

6. REFERENCES

[1] M. A. Hearst. Texttiling: segmenting text into multi-paragraph subtopic passages. *Comput. Linguist.*, 23(1):33–64, 1997.

[2] R. Yan, Y. Li, Y. Zhang, and X. Li. Event recognition from news webpages through latent ingredients extraction. In *AIRS '10*, pages 490–501.