Sentiment Classification in Resource-Scarce Languages by using Label Propagation

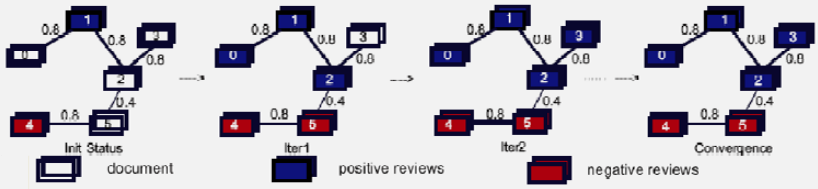
Kitsuregawa,Toyoda Lab.
renyong@tkl.iis.u-tokyo.ac.jp
Ren Yong

Research Background

Existing methods for sentiment classification: supervised classification
Main advantage: heavily depend on a large amount of labeled data (labor and time consuming)
Different approach (less requirement on the number of training data) is required in resource-scarce language (a sufficient amount of training data is not available)

Build Similarity Graph for Label Propagation

Label propagation: graph-based semi-supervised learning
Similarity graph is foremost factor
In ideal similarity graph, the more similar two vertices, the higher the similarity score between them



Various methods have been investigated to build better graph

Three kinds of sentiment features
1) Content words
2) Phrases (extracted by using tailored POS)
3) Adjectives

Different similarity measures

Name	Computing formula
Dice	$\frac{2\lvert A\cap B\rvert}{\lvert A\rvert+\lvert B\rvert}$
Overlap	$\frac{\lvert A\cap B\rvert}{min(\lvert A\rvert,\lvert B\rvert)}$
Jaccard Index	$\frac{\lvert A\cap B\rvert}{\lvert A\cup B\rvert}$
Cosine (binary)	$\frac{\lvert A\cap B\rvert}{\sqrt{(\lvert A\rvert\times\lvert B\rvert)}}$
Cosine (tf-idf)	$\frac{A\cdot B}{\lvert A\rvert\lvert B\rvert}$

Evaluation

ChnSentiCorp[1] includes Chinese reviews from three domains, each domain is about 4000 reviews
Data splitting: test (300), training data includes labeled seeds (10~300) and unlabeled data
Performance measure: ten-fold experiment, average accuracy

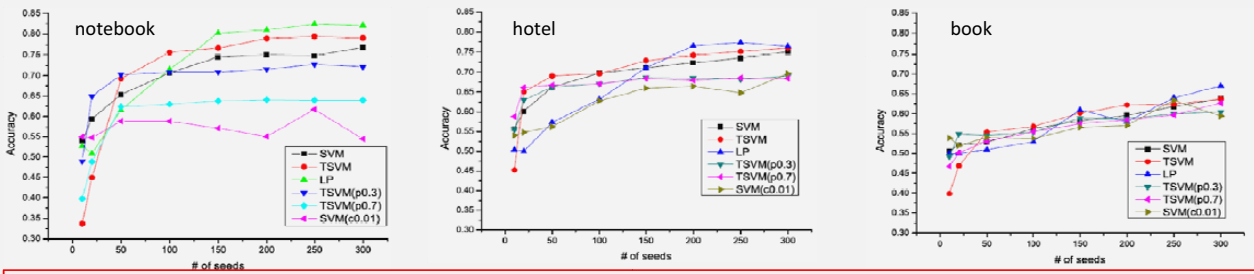
Comparison of sentiment features

Domains	Sentiment features	Cosine (tf-idf)	Cosine (binary)	Dice	Jaccard	Overlap
Notebook	Phrases	0.826	0.821	0.820	0.819	0.819
	Content words	0.584	0.681	0.69	0.685	0.611
	Adjective words	0.59	0.509	0.514	0.52	0.507
Hotel	Phrases	0.752	0.764	0.712	0.759	0.735
	Content words	0.528	0.532	0.544	0.534	0.502
	Adjective words	0.507	0.503	0.501	0.501	0.501
Book	Phrases	0.598	0.631	0.641	0.623	0.627
	Content words	0.626	0.619	0.626	0.604	0.535
	Adjective words	0.507	0.503	0.501	0.501	0.501

Comparison of similarity measures

Domain	Cosine (tf-idf)	Cosine (binary)	Dice	Jaccard	Overlap
Notebook	0.826	0.821	0.820	0.819	0.819
Hotel	0.752	0.764	0.712	0.759	0.735
Book	0.598	0.631	0.641	0.623	0.627

Performance comparison with SVM and TSVM



1.Phrases perform best as sentiment features
2.Classification performance is not greatly affected by the choice of similarity measures
3.LP is parameter-free and stable
4.When appropriate seeds are available, LP could outperform SVM and TSVM

[1] http://www.searchforum.org.cn/tansongbo/corpus-senti.htm