Секция «Вычислительная математика и кибернетика»

Метод отбора N-грамм для формирования лексикона коллекции текстовых документов

Царьков Сергей Валерьевич
Аспирант
Московский физико-технический институт, факультет инноваций и высоких технологий, Москва, Россия
E-mail: s.v.tsarkov@gmail.com

Для тематического моделирования большой коллекции документов необходимо сформировать словарь ключевых фраз, которые могли бы характеризовать тематику документов. Процесс выделения ключевых фраз включает этап ранжирования N-грамм — фраз из N слов — с последующим составлением списка из k фраз, наиболее «весомых» по некоторому критерию [2], например *C-value* [1].

С помощью морфологического словаря и базы правил образования словосочетаний из текста выделяются только те N-граммы, между словами которых существуют синтаксические связи. Для построения N-грамм производится объединение $(N-1)$-грамм с синтаксически связанными с ними 2-граммами. В результате строится лексикон коллекции документов — словарь фраз, хранящий также их статистические характеристики. Эксперименты показывают, что объем лексикона быстро растёт по мере добавления документов, главным образом за счет фраз, не являющихся ключевыми. Для сокращения лексикона предлагается отбрасывать фразы, встречающиеся в документе менее t раз. Так как фраза не может встречаться чаще, чем её часть, для построения N-грамм можно использовать только $(N-1)$-граммы, встречающиеся в документе не менее t раз, что позволяет повысить скорость построения лексикона.

Целью работы является исследование влияния параметра t на объём и качество лексикона. Эксперименты проводились на коллекции из 2000 русскоязычных документов. Было получено 3 лексикона при $t = 1, 2, 3$. Для каждого лексикона и каждого документа был составлен список $k = 100$ лучших N-грамм, ранжированных по критерию *C-value*. Точность списка b оценивалась как доля его фраз, входящих в эталонный список e, полученный при $t = 1$. Кроме того, оценивалась средняя ошибка ранжирования списка b:

$$PE(b) = \frac{1}{|b|} \sum_{s \in b} |p(s, e) - p(s, b)|,$$

где $p(s, b)$ — позиция фразы s в списке b. Общая ошибка ранжирования PE лексикона оценивалась как среднее PE по всем документам.

При $t = 1$ объём лексикона 5 066 852 фраз. При $t = 2$ объём сокращается в 5 раз (904 588), при этом точность 81%, $PE = 10{,}56$. При $t = 3$ объём 352 920, точность 68%, $PE = 33{,}15$.

Работа выполнена при поддержке Министерства образования и науки РФ в рамках Государственного контракта 07.524.11.4002.

Литература

1. Frantzi K., Ananiadou S., Mima H. Automatic recognition of multi-word terms: the C-value/NC-value method // Intl. J. of Digital Libraries Vol. 3 Issue 2, 2000, p. 117-132.

2. Hussey R., Williams S., Mitchell R. Automatic keyphrase extraction: a comparison of methods // In Proc. of the 4th International Conference on Information, Process, and Knowledge Management, 2012, p. 18-23.