О выборе модели представления текстовой информации для задач анализа и фильтрации содержимого Интернет трафика

Петровский М. И., Глазкова В. В., Царёв Д. В.
Москва, МГУ им. М.В.Ломоносова

Задача классификации текстовых и гипертекстовых документов — одна из основных алгоритмических подзадач, возникающих при реализации систем фильтрации Интернет трафика, в частности, систем фильтрации спама и web-контента. Объекты классификации — текстовые и гипертекстовые документы и их фрагменты — являются слабо структурированными разнородными данными. Большинство алгоритмов классификации работают с формальным описанием объектов, используя векторную модель представления документа [1]. В данной модели документ описывается числовым вектором фиксированной длины $\bar{a} \in \mathrm{R}^n$, где размерность вектора n есть число признаков, а i-я координата определяет вес i-ого признака. Соответственно, для реализации модели представления необходимо, во-первых, выбрать признаковое пространство, во-вторых, определить алгоритм вычисления весов. Качество выбранной модели представления при фиксированном алгоритме классификации и фиксированном тестовом наборе документов можно оценить по следующим критериям: точность классификации, размерность признакового пространства, размер получаемой модели классификации, время обучения и классификации, поддержка морфологии.

Самым распространенным способом формирования признакового пространства является *метод ключевых слов*, где в качестве признаков используются лексемы, входящие в документы, а размерность пространства равна размеру словаря. Но данный метод не учитывает морфологию языка. Поддержку морфологии можно реализовать с помощью stemming (все слова приводятся к своим базовым словоформам), что приводит к дополнительной вычислительной нагрузке. Кроме того, построение лексического анализатора для некоторых языков является достаточно сложной задачей. Более просто проблему морфологии решает разбиение лексем на N-граммы. В этом случае в качестве координат в признаковом пространстве рассматриваются все возможные подряд идущие буквосочетания фиксированной длины N. При этом однокоренные слова образуют сходный набор N-грамм.

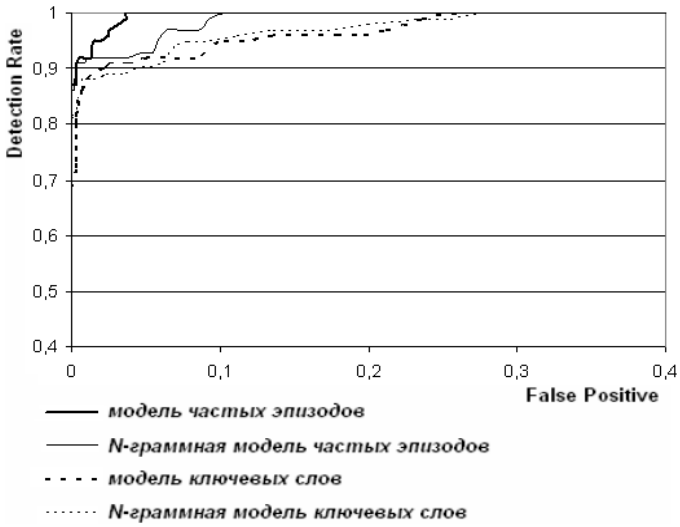
Необходимо отметить, что основным недостатком обоих подходов является то, что семантические связи между лексемами (или N-граммами) не учитываются. Для преодоления этого недостатка нами был предложен метод построения модели представления, основанный на выделении *частых эпизодов*. В этом случае множество частых комбинаций лексем

	лексемы	N-gramm	эпизоды	эпизоды+N-gramm
`Detection Rate`	89%	85%	**92**%	91%
`False Positive`	1.2%	0.5%	0.5%	**0.3**%
`Размерность`	4227	5042	**3331**	5172
`Размер модели`	712	698	**610**	612

(или N-грамм) формирует новое признаковое пространство. Для этого на этапе обучения из каждого документа выделяются все предложения, входящие в его состав. Каждое предложение представляет собой отдельную транзакцию t, состоящую из лексем (или N-грамм) данного предложения. Весь тренировочный набор документов представляется в виде множества таких транзакций $\{t\}$. Далее с помощью алгоритма FP-tree [2] в $\{t\}$, выделяются частые эпизоды лексем (или N-грамм), которые удовлетворяют заданным параметрам — минимальной частоте встречаемости и максимальному размеру эпизода. Все полученные эпизоды нумеруются и составляют новое признаковое пространство.

Вес i-го признака определятся как нормированная частота встречаемости этого признака в документе: $a_i = f_i \,/\, \sqrt{\sum_{i=1}^{m} f_i^2}$, где f_i — частота встречаемости i-го признака в документе, m — количество непустых признаков в данном документе. В отличие от традиционных мер сходства типа TF-IDF, предложенную меру можно использовать при дообучении, не пересчитывая веса, поскольку они зависят только от текущего документа, а не от всего набора.

Для сравнения моделей представления нами был использован эталонный тестовый набор документов SpamAssassin public corpus [3], который содержит как текстовые, так и гипертекстовые документы, относящиеся к одному из двух классов: спам и легальная почта. В качестве базового алгоритма классификации был выбран алгоритм на основе метода опорных векторов (SVM), как один из наиболее популярных и точных алгоритмов для классификации данных большой размерности. В качестве меры сходства мы использовали экспоненциальную потенциальную функцию. Результаты экспериментов представлены ниже в виде ROC-кривых для сравнения точности и сводной таблицы по всем основным критериям. Размер модели указан как число опорных векторов в построенной SVM модели. Время обучения для методов на основе частых эпизодов существенно больше, но для задачи фильтрации Интернет трафика это не принципиально, поскольку обучение может производиться в offline. Время классификации у всех алгоритмов получилось примерно одинаковым.



Из результатов эксперимента видно, что предложенная модель на основе частых эпизодов кардинально превосходит традиционные подходы по всем основным критериям. Кроме того, представление на основе частых эпизодов с N-граммами можно использовать для морфологически сложных языков, таких как русский и немецкий.

Работа выполнена при поддержке гранта РФФИ №06-01-00691, гранта Президента РФ МК-4264.2007.9, а также в рамках госконтракта с Федеральным агентством по науке и инновациям №02.514.11.4026.

Литература

[1] *Salton G., McGill J.* An introduction to modern information retrieval. — New York: McGraw-Hill, 1983.

[2] *Jian Pei* Pattern-growth Methods for Frequent Pattern Mining. — Ph.D. Thesis, Simon fraser university, 2002.

[3] Apache Software Foundation. — The Apache SpamAssassin Public Corpus. — `http://spamassassin.apache.org/publiccorpus/`.