Выделение фрагментов в текстах при классификации

Markup of text fragments during classification

Васильев В. Г. (vvg_2000@mail.ru)

Институт проблем информатики РАН

В работе проводится сравнительный анализ подходов к выделению значимых фрагментов в текстах в процессе автоматической классификации. Рассматриваются новые алгоритмы на основе скрытой марковской модели, покрытия текста специальным иерархическим множеством фрагментов и предварительной сегментации текстов.

1. Введение

При решении задач автоматической классификации текстовых данных одной из важных задач является представление и объяснение результатов классификации пользователю. В частности, достаточно важным для понимания причин отнесения текста к определенной рубрике является выделение в нем релевантных ей фрагментов (особенно это актуально в случае классификации политематических документов).

В случае использования лингвистического подхода (т. е. правила отнесения текстов к рубрике описываются с помощью некоторого информационно-поискового языка) такое выделение является достаточно простой задачей, которая решается путем отбора предложений, удовлетворяющих введенному правилу. Однако при использовании статистического подхода к классификации ее решение значительно усложняется. Это связано с тем, что в данном случае документ обычно представляется в виде одного вектора весов информационных признаков для всего текста, а также с отсутствием обучающих массивов с эталонным делением текстов на фрагменты.

Введем необходимые обозначения. Пусть $\omega_1, ..., \omega_k$ рубрики иерархического классификатора, задающие темы, которые представляют интерес и которые требуется автоматически выделять в текстах, $X = (X_1, ..., X_n)$ — текст на естественном языке, состоящий из *n* предложений, X_i — вектор весов слов в предложении $i = 1, ..., n$ размерности *m*, где *m* — общее число слов в тексте. Требуется для каждой рубрики ω_j, $j = 1, ..., k$, определить факт наличия в тексте информации по ней и в случае положительного решения найти предложения, ей соответствующие. Иными словами, для рубрики ω_j, $j = 1, ..., k$, требуется найти вектор: $t_j = (t_{j1}, ..., t_{jn})$, где

$$t_{ji} = \begin{cases} 1, & X_i \in \omega_j, \\ 0, & X_i \notin \omega_j. \end{cases}$$

Задача выделения значимых фрагментов тесно связана со следующими классическими задачами анализа текстов:

- классификация текстов — определение принадлежности текстов к рубрикам классификатора (в данном случае производится оценка отдельных предложений);
- сегментация текстов — разделение текстов на тематически однородные фрагменты (в данном случае производится выделение в тексте групп предложений, относящихся к одной теме);
- реферирование текстов — выделение значимых предложений в тексте с целью построения его краткого изложения (в данном случае производится выделение не всех значимых предложений, а только относящихся к определенной тематике).

Основной сложностью при выделении значимых фрагментов в текстах является то, что в общем случае оценку принадлежности предложения к рубрике нельзя проводить без учета соседних предложений. Например, возможна ситуация, когда фрагмент текста X, состоящий из предложений $X_i, X_{i+1}, ..., X_{i+s}$, относящийся к рубрике

ω_j, $j = 1, ..., k$, при классификации целиком будет отнесен к данной рубрике, а при классификации предложений $X_i, X_{i+1}, ..., X_{i+s}$ по отдельности ни одно из них может быть не отнесено к данной рубрике ω_j. Также возможна и обратная ситуация, что предложение X_i относится к рубрике ω_j, но при рассмотрении вместе с соседними предложениями оно уже не будет отнесено к данной рубрике.

Рассмотрим более подробно методы сегментации текстов, которые являются наиболее близкими по своему содержанию к задаче, решаемой в настоящей работе. Можно выделить следующие основные подходы к их построению:

- процедурный подход;
- структурный подход;
- вероятностный подход;
- оптимизационный подход.

Процедурный подход основан на построении правил, учитывающих различные элементы текста: отступы строк, знаки препинания, ключевые слова, референтные связи между словами, а также различные элементы оформления документов (заголовки, разделы, параграфы). Данный подход оказывается эффективным только в том случае, если формат обрабатываемых документов является известным.

Структурный подход основан на использовании различных мер близости между предложениями или фрагментами текста. При этом возможно как вычисление простейших статистик совместной встречаемости слов в различных блоках текста, так и использование методов кластерного анализа. При этом наибольшее распространение получил метод скользящего окна или «перекрывающегося текста» (text-tiling) [2], основанный на нахождении мест в тексте, где мера близости между двумя соседними блоками предложений минимальна. Также интересным является подход, основанный на использовании дивизимного алгоритма иерархического кластерного анализа [3].

Вероятностный подход для сегментации текстов основан на построении различных вероятностных моделей порождения слов в текстах. На практике наибольшее распространение получило представление текстов с помощью скрытых марковских моделей [5, 1]. В частности, в работе [5] рассматривается задача разделения составного текста, представляющего запись передач новостей по радио, на отдельные новостные сообщения. В данном случае открытым состояниям скрытой марковской модели, формальное определение которой будет дано далее, соответствуют отдельные слова в тексте, а скрытым состояниям — позиции слов в отдельных сообщениях, т. е. первым словам каждого сообщения будет соответствовать состояние с номером 1.

Оптимизационный подход основан на задании некоторого показателя качества разделения текста на фрагменты и нахождении такого разделения, которое обеспечивало бы его максимум. В частности, в работе [4] показатель качества зависит от длины фрагмента текстов и степени близости соседних фрагментов текстов. Для нахождения максимума показателя используется алгоритм на основе методов динамического программирования.

Недостатком приведенных подходов к сегментации является то, что при разделении текста на фрагменты в них не учитывается известная информация о тематиках, которые интересуют пользователя (в нашем случае его интересы заданы в виде набора рубрик классификатора). Рассмотрим несколько возможных подходов к решению задачи разметки текстов, которые, с одной стороны, используют идеи, приведенных выше подходов к сегментации текстов, а с другой стороны, лишены указанного недостатка и опираются на использование ранее обученных классификаторов.

2. Выделение фрагментов путем построения иерархического покрытия текста

Пусть текст $X = (X_1, ..., X_n)$ представляется в виде множества векторов

$$F = \left\{ \sum_{t=l_1}^{l_2} X_t \mid 1 \le l_1 \le l_2 \le n \right\},$$

где X_i — вектор весов слов (словосочетаний) предложения $i = 1,...,n$, n — число предложений в тексте. Данное множество включает в себя множество всех непрерывных фрагментов (последовательностей предложений без пропусков), содержащихся в тексте. Для оценки степени соответствия предложения текста рубрике можно использовать следующие выражения:

$$w_{ij} = \max_{Y \in F, X_i \in Y} g_j(Y) \text{ или}$$

$$w_{ij} = \underset{Y \in F, X_i \in Y}{mean}\, g_j(Y),\ i = 1, ..., n,\ \ j = 1, ..., k,$$

где $g_j(Y)$ — функция, осуществляющая вычисление степени соответствия вектора Y рубрике ω_j, $j = 1, ..., k$, построенная в результате обучения некоторого статистического классификатора.

Таким образом, вес w_{ij} предложения X_i, $i = 1, ..., n$, для рубрики ω_j равен максимальному (среднему) значению степени близости к данной рубрике ω_j всех фрагментов, содержащих данное предложение. Несложно заметить, что в данном случае для выделения значимых фрагментов в тексте для каждой рубрики требуется выполнить классификацию $\frac{n \cdot (n+1)}{2}$ фрагментов, т. е. вычислительная сложность составляет порядка $O(n^2)$.

Для снижения вычислительной сложности можно воспользоваться представлением текста X в виде следующего иерархически упорядоченного множества векторов фрагментов (иерархического покрытия):

$$H = H_0 \cup \left(\bigcup_{t=1}^{\lceil \log_2(n) \rceil} H_t \right),$$

$$H_t = \left\{ \sum_{i=1+l2^{t-1}}^{\min\left(l2^{t-1}+2^t, n\right)} X_i \mid l = 0,..., \left\lceil \frac{n}{2^{t-1}} - 1 \right\rceil \right\},$$

$$H_0 = \{X_1, ..., X_n\},$$

где $\lceil x \rceil = \min(l \in Z \mid l \geq x)$.

Несложно заметить, что для мощности множества H справедливы следующие соотношения:

$$|H| = |H_0| + \sum_{t=1}^{\lceil \log_2(n) \rceil} |H_t| \leq n + \sum_{t=1}^{\lceil \log_2(n) \rceil} \left(\frac{n}{2^{t-1}} + 1 \right) \leq$$

$$\leq \log_2(n) + n + 1 + n \sum_{t=1}^{\infty} \frac{1}{2^{t-1}} \leq \log_2(n) + 1 + 3n$$

Таким образом, при использовании иерархического покрытия H вычислительная сложность нахождения степени соответствия текста отдельной рубрике составляет порядка $O(n)$.

Для построенного иерархического покрытия H справедлива следующая теорема, которая говорит о качестве аппроксимации полного множества фрагментов F с помощью множества H.

Теорема. (Об иерархическом покрытии.)

Для любого фрагмента $Y \in F$ существует фрагмент $Z \in H$ такой, что

$$\frac{|Y \Delta Z|}{|Y|} \leq \frac{1}{2}.$$

Доказательство. Рассмотрим следующие три случая для числа предложений во множестве $Y \in F$.

1. Пусть $|Y| = 2^l$, $l \in \{0, 1, ..., \lceil \log_2(n) \rceil\}$.

Тогда $Y = \sum_{i \in [s, s+2^l-1]} X_i$, где $s \in \{1, ..., n - 2^l + 1\}$,

и существует $Z = \sum_{i=1+v2^{l-1}}^{\min\left(v2^{l-1}+2^l, n\right)} X_i \in H$ такое,

что $\left| s - \left(1 + v2^{l-1}\right) \right| \leq 2^{l-2}$,

v — неотрицательное целое, т. е. начало фрагмента соответствующего Z расположено не далее чем в 2^{l-2} предложениях от начала фрагмента Y.

Отсюда получаем, что $|Y \Delta Z| \leq 2 \cdot 2^{l-2} = \frac{1}{2}|Y|$.

2. Пусть $|Y| \geq 2^l + 2^{l-1}$ и $|Y| < 2^{l+1}$.

$$Y = \sum_{i \in [s, s+2^l+2^{l-1}+\delta-1]} X_i \quad , \text{где}$$

$$s \in \{1, ..., n - 2^l - 2^{l-1} - \delta + 1\},$$

$$\delta \in [1, ..., 2^{l-1} - 1],$$

и существует $Z = \sum_{i=1+v2^{l-1}}^{\min\left(v2^{l-1}+2^l, n\right)} X_i \in H$ такое,

что $Z \subset Y$, v — неотрицательное целое.

Отсюда получаем, что

$$|Y \Delta Z| = |Y \setminus Z| \leq 2^{l-1} + 2^{l-1} = \frac{1}{2}|Y|.$$

3. Пусть $|Y| < 2^l + 2^{l-1}$ и $|Y| > 2^l$.

Тогда $Y = \sum_{i \in [s, s+2^l+\delta-1]} X_i$,

где $s \in \{1, ..., n - 2^l - \delta + 1\}$,

и существует $Z = \sum_{i=1+v2^{l-1}}^{\min\left(v2^{l-1}+2^l, n\right)} X_i \in H$ такое,

что $\left| s - \left(1 + v2^{l-1}\right) \right| \leq 2^{l-2}$, v — неотрицательное целое.

Отсюда получаем, что

$$|Y \Delta Z| \leq 2 \cdot 2^{l-2} - \delta \leq \frac{1}{2}|Y|. \blacksquare$$

Таким образом, для каждого фрагмента Y из F существует фрагмент из H, который отличаются от него не более чем на половину его длины. В целом схема алгоритма классификации текста принимает следующий вид.

3. Схема работы алгоритма на основе иерархического покрытия

1. Построение иерархического покрытия множества предложений текста.
2. Выполнение независимой классификации фрагментов текста с использованием ранее обученного классификатора.
3. Вычисление итоговых весов предложений в тексте путем объединения результатов классификации фрагментов, входящих в покрытие.
4. Отбор предложений, вес которых выше некоторого порога. ■

4. Выделение фрагментов с использованием скрытой марковской модели

В соответствии с работой [6] скрытая марковская модель с дискретным временем определяется как набор следующих элементов:

$S = \{s_1, ..., s_N\}$ — множество скрытых состояний;

$q_1, ..., q_n \in S$ — последовательность скрытых состояний;

A — матрица переходных вероятностей размера $N \times N$, где $a_{ij} = p(q_t = s_j \mid q_{t-1} = s_i)$.

U — пространство наблюдаемых состояний;

$y_1, ..., y_n \in U$ — последовательность наблюдаемых состояний;

$f_1(u), ..., f_N(u)$ — условные функции распределения для состояний s_i, $i = 1, ..., N$;

$\pi = (\pi_1, ..., \pi_N)$ — вектор начальных вероятностей скрытых состояний.

Рассмотрим теперь как с использованием аппарата скрытых марковских моделей можно производить выделение в тексте X предложений, соответствующих отдельной рубрике ω_r, $r = 1, ..., k$.

В данном случае пространство наблюдаемых состояний $U = R^m$, где m — размерность словаря признаков для обучающей выборки, а элементы последовательности наблюдаемых состояний $y_i = X_i$, $i = 1, ..., n$.

Множество скрытых состояния $s = \{s_1, s_2, s_3, s_4\}$ определим следующим образом:

s_1 — предложение находится внутри фрагмента, соответствующего рубрике w_r;

s_2 — предложение находится в начале фрагмента, соответствующего рубрике w_r;

s_3 — предложение находится в конце фрагмента, соответствующего рубрике w_r;

s_4 — предложение не относится к рубрике ω_r.

Так на этапе обучения эталонное распределение текста на фрагменты отсутствует, то матрицу переходных вероятностей A зададим априори. В данной работе при проведении экспериментов использовалась следующая матрица

$$A = \begin{pmatrix} 0.8 & 0 & 0.2 & 0 \\ 0.8 & 0.2 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 \\ 0 & 0.5 & 0 & 0.5 \end{pmatrix}.$$

Начальные вероятности скрытых состояний положим равными друг другу, т. е. $\pi_j = 1/4$, $j = 1,...,4$.

Основную сложность при построении данной модели составляет задание условных функций распределения (плотности) $f_j(u)$ для состояний s_j, $j = 1,...,4$,
и вычисление их значений $w_{ij} = f_j(X_i)$.

В настоящей работе значения данных функций определяются следующим образом. Каждому предложению X_i ставятся в соответствие следующие вектора:

$$y_i = \sum_{t=-b}^{b} w_t X_{i+t},\ y_i^- = \sum_{t=-b}^{0} w_t X_{i+t},\ y_i^+ = \sum_{t=0}^{b} w_t X_{i+t}$$

где b — константа, задающая размер блока (число учитываемых предложений справа и слева от предложения с номером i), w_t — вес предложения с индексом $i + t$, $i = 1, ..., n$,

$w_t = (1 + |t|)^{-1}$, $w_t = 1$, t — целое.

Отсюда w_{ij}, $i = 1,...,n$, определяются следующим образом:

$$w_{i1} = p(y_i \mid \mathsf{w}_r),$$
$$w_{i2} = p(y_i^+ \mid \mathsf{w}_r),$$
$$w_{i3} = p(y_i^- \mid \mathsf{w}_r),$$
$$w_{i4} = (1 - p(y_i \mid \mathsf{w}_r)).$$

Для определения последовательности скрытых состояний используется стандартный алгоритм динамического программирования (алгоритм Витерби [6]), который находит наиболее правдоподобную последовательность из всех возможных. Таким образом, Общая схема работы алгоритма выделения и классификации фрагментов в результате приобретает следующий вид.

5. Схема работы алгоритма на основе скрытой марковской модели

1. Выделить предложения в тексте и сопоставить им векторы информационных признаков.
2. Произвести классификацию отдельных предложений с помощью обученного ранее статистического классификатора.
3. С использованием алгоритма Витерби [6] произвести оценивание скрытых состояний цепи Маркова по наблюдаемым состояниям.
4. Произвести разметку текста с использованием найденных скрытых состояний. ■

6. Выделение фрагментов путем независимой сегментации

В данном случае сначала выделение фрагментов производится путем сегментации текста на тематически однородные фрагменты без учета информации о структуре рубрик классификатора. Затем проводится классификация выделенных фрагментов и отбираются те из них, которые удовлетворяют рубрикам классификатора. В настоящей работе для сегментации текстов было решено остановится на методах, описанных в работах [2] и [3].

В первом методе, который называется Text Tiling, сегментация текста проводится следующим образом. Для каждого промежутка между предложениями вычисляется косинусная мера близости блока из r предложений справа и слева от него. В результате формируется вектор s размерности n, где n — число предложений, r — размер блока (в экспериментах использовалось значение $r = 4$). Далее значения вектора s сглаживаются с использованием метода скользящего среднего с различными параметрами и находятся точки локального минимума, которые и используются в качестве границ фрагментов.

Второй метод основан на использовании иерархического алгоритма кластерного анализа, который последовательно формирует разбиения данных на кластеры, начиная с ситуации, когда в одном кластере содержаться все наблюдения, и заканчивая ситуацией, когда каждое наблюдение образует отдельный кластер. Схема вычислений в данном случае следующая.

7. Схема алгоритма сегментации на основе иерархического кластерного анализа

1. Вычисляется матрица $C = (c_{ij})$ размерности $n \times n$,

элементы которой $c_{ij} = \dfrac{X_i^T X_j}{\|X_i\|_2 \|X_j\|_2}$,

$i, j = 1,\ldots,n$.

2. Строится матрица локальных весов $R = (r_{ij})$ размерности $n \times n$, где

$$r_{ij} = \frac{\left|\forall p \in [1, i-r, i+r], q \in [j-r, j+r], p, q \in 1,\ldots,n : m_{ij} > m_{pq}\right|}{(2r+1)^2}$$

$$i, j = 1,\ldots,n,$$

r — константа, задающая размер окна, в рамках которого вычисляются локальные веса.

Осуществляется рекурсивное разбиение существующих фрагментов на части, начиная с ситуации, когда все предложения относятся к одному фрагменту, и, заканчивая ситуацией, когда перестанет возрастать функция среднего значения внутрифрагментной близости μ_T, которая определяется следующим образом

$$\mu(T) = \frac{\sum_{k=1}^{|T|} \sum_{i \in t_k} \sum_{j \in t_k} r_{ij}}{\sum_{k=1}^{|T|} |t_k|^2}, \text{ где}$$

$T = \{t_1,\ldots,t_{|T|}\}$ — множество фрагментов в тексте. ■

В целом схема работы алгоритма классификации фрагментов текста в данном случае принимает следующий вид.

8. Схема работы алгоритма на основе независимой сегментации

1. Выделить предложения в тексте и сопоставить им векторы информационных признаков.
2. Произвести разбиение предложений текста на непересекающиеся фрагменты с использованием алгоритма сегментации.
3. Выполнить классификацию выделенных фрагментов с использованием обученного ранее классификатора и отобрать фрагменты, удовлетворяющие хотя бы одной рубрике классификатора.
4. Произвести разметку текста по результатам классификации фрагментов. ■

9. Экспериментальная оценка эффективности подходов к выделению фрагментов

Для оценки влияния процедур выделения фрагментов на итоговое качество классификации были проведены два эксперимента с массивом нормативно-правовых документов, используемом в рамках семинара РОМИП в 2004 году.

10. Эксперимент по оценке качества классификации с учетом деления на фрагменты

В первом эксперименте из-за отсутствия эталонного массива с размеченными фрагментами производилась оценка качества классификации текстов целиком с учетом выделения фрагментов различными алгоритмами. Для обучения и оценивания

использовались документы, входящие обучающее множество коллекции РОМИП-2004. Данное множество было преобразовано следующим образом. Сначала были отобраны 44 рубрики, содержащие не менее 50 документов. Затем множество отобранных документов было разбито в пропорции 80% на 20% для построения нового обучающего и тестового множеств.

Для классификации текстов использовались метод вероятностной классификации на основе смеси распределений фон Мизеса-Фишера (VMF) и метод на основе машин опорных векторов (SVM). Фрагменты размером менее 4 предложений при классификации не учитывались. Множество рубрик, к которым относится текст целиком, формировалось путем объединения множеств рубрик, к которым были отнесены отдельные фрагменты. Таким образом, в результате классификации текста ему ставится некоторое множество рубрик. Оценка качества в данном случае производится с использованием стандартных коэффициентов точности, полноты и F-меры с использованием микро усреднения.

Результаты оценки качества классификации приведены в следующей таблице. В ней используются следующие обозначения: NONE — классификация без выделения фрагментов; HIER — выделение фрагментов на основе построения иерархического покрытия; HMM — выделение в тексте фрагментов с использованием скрытой марковской модели; TILE — выделение фрагментов путем предварительной сегментации текста с помощью алгоритма Text Tile; DIV — выделение фрагментов путем предварительной сегментации текста с использованием алгоритма дивизимного кластерного анализа. Для каждого метода через запятую приводятся показатели: P — точность, R — полнота, F — F-мера.

Таблица 1. Качество классификации с выделением фрагментов

Метод	SVM (P, R, F)	VMF (P, R, F)
NONE	36%, 60%, 43%	22%, 78%, 32%
HIER	29%, 70%, 38%,	15%, 85%, 24%
HMM	26%, 67%, 36%	9%, 91%, 16%
TILE	30%, 61%, 40%	15%, 83%, 25%
DIV	34%, 63%, 41%	16%, 81%, 26%

Анализ результатов эксперимента, приведенного в таблице 1 позволяет сделать вывод, что использование практически всех методов выделения фрагментов приводит к повышению полноты классификации, что является следствием рассмотрения большего количества элементов текста. При этом наиболее повышение полноты достигается в случае использовании метода HIER, а наименьшее при использовании метода TILE.

11. Эксперимент по оценке качества выделения фрагментов в текстах

Во втором эксперименте для оценки качества выделения фрагментов было сформировано искусственное тестовое множество $\widetilde{T} = \{\widetilde{T}_1,...,\widetilde{T}_N\}$, элементы которого $\widetilde{T}_j = [T_{j_1}, T_{j_2},..., T_{j_h}]$ получаются в результате конкатенации случайного набора из h текстов из множества T. В данном случае N — число текстов во множестве $\widetilde{T}$, $j_1,..,j_h$ — индексы случайно выбранных с возвращением текстов из множества T, T — множество таких текстов из тестового множества, используемого в первом эксперименте, которые относятся только к одной рубрике.

Каждому тексту $\widetilde{T}_j = [T_{j_1}, T_{j_2},..., T_{j_h}]$, $j = 1,..., N$, была поставлена в соответствие матрица эталонной классификации предложений

$$C_j^e = \left(c_{jki}^e\right)$$

и матрица автоматической классификации

$$C_j^a = \left(c_{jki}^a\right),$$

где $c_{jki}^a, c_{jki}^e \in \{0,1\}$ — признак принадлежности предложения $i = 1,..., n_j$ к рубрике ω_r, $r = 1,..., k$, k — число классов. Необходимо отметить, что при формировании данной матрицы предполагалось, что все предложения из текстов T_{j_l}, $l = 1,..., h$, относятся к одному классу.

Для оценки качества выделения фрагментов использовались следующие показатели:

$$P_{jr} = \sum_{i=1}^{n_j} c_{jki}^e c_{jki}^a \Big/ \sum_{i=1}^{n_j} c_{jki}^a$$

и $R_{jr} = \sum_{i=1}^{n_j} c_{jki}^e c_{jki}^a \Big/ \sum_{i=1}^{n_j} c_{jki}^e$ — точность и полнота классификации предложений в тексте $j = 1,..., N$ для класса $r = 1,..., k$;

$$P = \frac{1}{Nr}\sum_{r=1}^{k}\sum_{j=1}^{N} P_{jr} \text{ и } R = \frac{1}{Nr}\sum_{r=1}^{k}\sum_{j=1}^{N} R_{jr} —$$

средние значения точности и полноты классификации предложений по всем классам.

В следующей таблице приводятся результаты экспериментов по оценке качества выделении фрагментов с использованием двух методов классификации. При проведении эксперимента использовались $h = 5$ и $N = 100$.

Таблица 2. Качество выделения фрагментов в текстах

Метод	SVM (P, R, F)	VMF (P, R, F)
NONE	20%, 23%, 22%	15%, 46%, 22%
HIER	**62%**, **72%**, **66%**	22%, **72%**, 33%
HMM	46%, 4%, 7%	3%, 3%, 3%
TILE	52%, 57%, 54%	29%, **72%**, **41%**
DIV	51%, 59%, 54%	33%, 48%, 37%

Таким образом, можно сделать следующий вывод, что качество выделения фрагментов оказывается на достаточно высоком уровне, учитывая недостатки, которые присущи массиву текстов РОМИП. При этом наилучшие показатели качества выделения фрагментов достигаются при использовании метода на основе иерархического покрытия. При этом использование алгоритма SVM оказывается предпочтительнее.

12. Заключение

Таким образом, в данной работе рассмотрены подходы к разметке текстов по результатам автоматической классификации, основанные на построении специального иерархического покрытия документов фрагментами, использовании скрытых марковских моделей и сегментации текстов. Проведенные эксперименты показали, что за счет выделения фрагментов в текстах можно повысить полноту классификации.

Актуальной задачей для дальнейших исследований является проведение более подробных исследований по оценке качества определения границ фрагментов в тексте, а также проведение с другими алгоритмами сегментации текста на фрагменты.

Работа выполнена при поддержке гранта Президента Российской Федерации для государственной поддержки молодых российских ученых МК-12.2008.10.

Литература

1. *Blei D., Moreno P. J.* Topic Segmentation with an Aspect Hidden Markov Model // SIGIR'01, September 9–12, 2001, New Orleans, Louisiana, USA. — 6 p.
2. *Chual T., Chang S., Chaisorn L., Hsu W.* Story Boundary Detection in Large Broadcast News Video Archives — Techniques, Experience and Trends // MM'04, October 10–16, 2004, New York, USA. — pp. 656–659.
3. *Choi F., Wiemer-Hasting P., Moore J.* Latent semantic Analysis for Text Segmentation // Proceedings of NAACL'01, Pittsburgh, PA, 2001. — pp. 109–117.
4. *Fragkou P., Petridis V., Kehagias Ath.* Linear Text Segmentation using a Dynamic Programming Algorithm, 2003. — 8 p.
5. *Greiff W., Morgan A., Fish R., Richards M., Kundu A.* Fine-grained hidden markov modeling for broadcast-news story segmentation // Proceedings of the First international Conference on Human Language Technology Research (San Diego, March 18 — 21, 2001). Human Language Technology Conference. Association for Computational Linguistics, Morristown, NJ, 2001. — pp 1–5.
6. *Rabiner L.* A tutorial on hidden markov models and selected applications in speech recognition // Proc. IEEE, 77 (2), 1989. — pp. 257–286.