ПРИМЕНЕНИЕ ЧАСТОТНО-КОНТЕКСТНОЙ КЛАССИФИКАЦИИ ТЕКСТОВОЙ ИНФОРМАЦИИ ПРИ ВЫБОРЕ ТЕКСТОВ ДЛЯ ИЗУЧЕНИЯ

В.В. Потараев

Белорусский государственный университет информатики и радиоэлектроники, Минск, Беларусь, vic229@rambler.ru

Abstract. Today informational systems contain lots of textual information. When a student prepares for an exam, he usually finds several articles in the internet. Text classification can help to select only one file which includes most information; for example, to select the text which has content similar to most of other texts. One of quite effective classification methods is probability-and-context based classification.

В современных информационных системах накоплено огромное количество текстовой информации. При изучении конкретной дисциплины и подготовке к зачёту или экзамену может оказаться недостаточно материалов из методички или конспекта. В таком случае студенты зачастую прибегают не к рекомендованным книгам, а к информации из интернета, что позволяет сэкономить время. Чаще всего, по каждой из изучаемых тем студент может найти сразу несколько статей, которые могут показаться полезными.

Предположим, необходимо выбрать один наиболее содержательный текст из множества текстов по данной вопросу. Очевидно, что в этом случае можно использовать классификацию текстовых данных. Классификация подразумевает вычисление близости текста с другими текстами (представляющими классы) [1]. Текст, который наиболее близок с остальными текстами (классами) по содержанию, будет наиболее полно отражать смысл всех найденных статей [2], и его следует рекомендовать для прочтения.

Для вычисления близости текстов можно сравнивать наборы высокочастотных (ключевых) слов в текстах – чем больше общих высокочастотных слов, тем тексты более близки. Если два текста (назовём их Текст1 и Текст2), содержащие определения некоторого термина, будут иметь единственное высокочастотное слово (этот термин), то и текст, который содержал бы определения из первых двух текстов (Текст3), имел бы то же самое ключевое слово. Для алгоритма выбора текста на основе частотной классификации такие тексты были бы эквивалентны, хотя Текст3 содержит наиболее полную информацию из трёх текстов.

Метод частотно-контекстной классификации использует графовую модель структурного представления текста произвольного содержания. Переход к модели структурного представления текста осуществляется описанным далее способом.

1) Текст рассматривается в виде информационного потока, образованного информационными элементами – словами.

Набор всех слов в тексте можно рассматривать как конечное множество уникальных информационных элементов: $I = \{i_1, i_2, \ldots, i_n\}$, где i – информационный элемент, соответствующий уникальному слову текста.

Информационный поток F, описывающий текст, будет представлен в виде набора этих элементов:

$F = (i_k, \ldots, i_m)$, $i_k, i_m \in I$, i_k соответствует первому, i_m – последнему слову в тексте.

Пример. Информационный поток, соответствующий заданному фрагменту текста: $F = (i_3, i_6, i_7, i_1, i_2, i_{11}, i_9, i_4, i_{10}, i_3, i_5, i_6, i_7, i_1, i_8, i_9, i_4, i_{10}, i_5)$.

2) Текст, представленный в виде информационного потока, формирует структуру.

Если учесть, что слова в тексте повторяются, то, соответственно, информационный поток будет многократно проходить через одни и те же информационные элементы, формируя ориентированный граф, вершинами которого являются слова, а ребрами – связи между словами в тексте.

Например, если информационный поток некоторого текста можно записать в виде $F = (i_3, i_6, i_7, i_1, i_2, i_{11}, i_9, i_4, i_{10}, i_3, i_5, i_6, i_7, i_1, i_8, i_9, i_4, i_{10}, i_5)$, то его информационная структура будет выглядеть следующим образом (рис. 1):

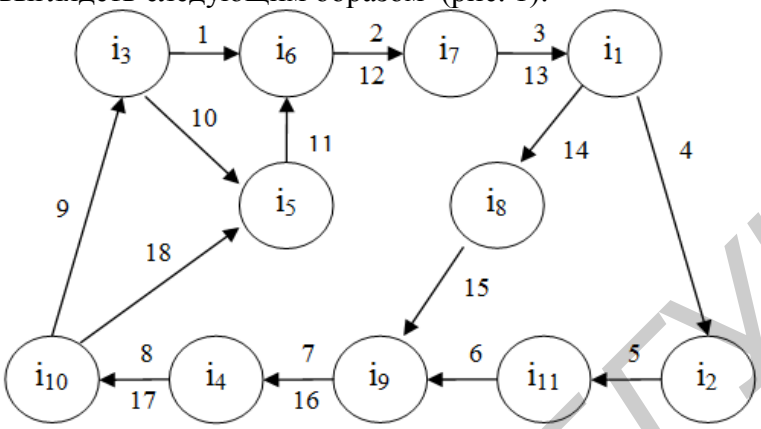

Рисунок. 1 – Структура, формируемая информационным потоком

Метод частотно-контекстной классификации текстовой информации основывается на гипотезе о том, что словарный запас и частоты использования слов зависят от темы текста. Тематическая классификация предполагает выделение множества ключевых слов, определяющих тематику текста. Для более точного выделения ключевых слов можно, помимо частоты встречаемости слов, учитывать и окружающие слова, то есть контекст.

Общая последовательность метода выглядит следующим образом [2]:

1) Моделирование текста и формирование его информационной структуры.

2) Выделение множества всех информационных элементов, ранжированных по их числу повторений в тексте.

3) Выделение множества ключевых элементов S_p.

4) Формирование уточняющего множества S_s на основе контекстного анализа информационных элементов множества S_p.

Если применить метод частотно-контекстной классификации к рассмотренному ранее примеру, то благодаря уточняющему множеству ключевых слов Текст3 является единственным текстом, который содержит все ключевые слова из двух других текстов. Поэтому Текст3 может быть выбран в качестве текста, рекомендуемого для изучения.

Таким образом, было доказано, что эффективность метода частотно-контекстной классификации выше, чем метода частотной классификации текстовой информации.

Литература

1. Когаловский, Р. Перспективные технологии информационных систем / Р. Когаловский М. – М.: ДМК Пресс, 2003. – 288 с.

2. Тарасов, С.Д. Метод тематического связанного ранжирования для автоматического сводного реферирования новостных сообщений в задачах поддержки принятия управленческих решений/ С.Д. Тарасов // Вестник ВГУ.– 2010. №1. – С. 166–173.