УДК 004.912

А. Ю. Селиванов, *Д. С. Андреев*[*]

ИЗВЛЕЧЕНИЕ ПЕРТИНЕНТНОЙ ИНФОРМАЦИИ ИЗ ТЕКСТОВЫХ ДОКУМЕНТОВ НА ОСНОВЕ БАЙЕСОВСКОГО КЛАССИФИКАТОРА С ИСПОЛЬЗОВАНИЕМ НЕЧЕТКИХ КОЛЛОКАЦИЙ

Извлечение информации или автореферирование – это процесс выделения смысловой, эмотивной, оценочной и иной информации из текстовых документов. На сегодняшний день каждому человеку доступны огромные объемы текстовой информации, которые невозможно обработать вручную. Вместе с тем, ежедневно появляются новые текстовые документы в любой из известных отраслей знаний.

Современный бизнес и производственную деятельность сложно себе представить без постоянного анализа новой актуальной информации. Для решения этой задачи часто привлекаются специальные аналитики – специалисты, профессионально занимающиеся поиском информации в определенных сферах. Результатом их работы являются реферативные отчеты о новых данных в исследуемой области. Автоматизация этого процесса позволит снизить издержки на получение новой пертинентной информации. Под пертинентностью понимается соответствие информации информационным интересам субъекта информационного поиска [1].

[*] Работа выполнена под руководством канд. техн. наук, доц. ФГБОУ ВО «ТГТУ» Д. В. Полякова.

Вместе с тем, есть ряд проблем, стоящий на пути автоматизации извлечения текстовой информации. Во-первых, невысокое качество моделей, анализирующих текстовые документы на основе семантики. Во вторых, тот факт, что искомой информацией часто является не весь текстовый документ, небольшой его фрагмент. Эта проблема особенно остро стоит при использовании статистических моделей. Небольшие текстовые фрагменты содержат мало термов, поэтому при вероятностной оценке их появления повышается роль статистических артефактов – термов случайно попавших во фрагмент, что приводит к ошибкам при их семантическом анализе.

Решением этой проблемы может стать использование коллокаций. Под коллокацией понимается коллективная локация термов, т.е. группа термов, расположенных рядом друг с другом [2]. Классически термы в коллокации располагаются непосредственно друг за другом. Однако есть работы [3], в которых термы могут располагаться на различном расстоянии друг относительно друга. Причем это расстояние может быть задано как некоторым целым числом, так нечетким числом. Коллокации, в которых расстояние между термами задано нечетким числом, получили название нечетких коллокаций [3]. Под расстоянием между термами понимается число других термов, расположенных в тексте между заданными, составляющими коллокацию. В некоторых работах [3] введено не только определение нечеткой коллокации, но приведен подход к выявлению семантически значимых коллокаций и построению функции принадлежности, формализующих расстояние между термами в коллокациях.

Учет нечетких коллокаций позволяет увеличить количество элементов, на основе которых исследуется текстовый фрагмент. Пусть n – число элементов в текстовом фрагменте. Рассмотрим только биграммы – коллокации состоящие из двух термов. В этом случае количество (N_2) элементов учтенных при формализации текстового фрагмента возрастет до

$$N_2 = n + C_n^2. \qquad (1)$$

Приняв во внимание равенство $N_1 = n = C_n^1$, получим, что (1) удобно обобщить для случая формализации текстового документа с учетом коллокаций состоящих из i термов, где $i = \overline{1,k}$, $k \le n$.

$$N_k = \sum_{i=1}^{k} C_n^i. \qquad (2)$$

Заметим, что количество термов в коллокации не может превосходить общее число термов текстового документа, а значит ограничение $k \le n$ естественно.

Из (2) легко видеть, что при учете всех возможных коллокаций количество элементов формализованного текстового документа возрастает экспоненциально относительно числа термов в рассматриваемом документе. Действительно, в силу (2)

$$N_n = \sum_{i=1}^{n} C_n^i = 2^n. \qquad (3)$$

Выражения (2) и (3) свидетельствуют о том, что учет коллокаций, в том числе нечетких, позволяет с уверенностью заявить о возможности кардинального увеличения элементов, формализующих текстовый документ. Это, в свою очередь говорит о целесообразности использования нечетких коллокаций при формализации небольших текстовых фрагментов. Назовем такие элементы, в число которых могут входить как термы, так и нечеткие коллокации, характеристическими элементами текстовых документов.

Для того, чтобы собрать пертинентную информацию, необходимо формализовать информационные интересы субъекта информационного поиска. Для этого введем в рассмотрение некоторое множество $D_{benefit}$ текстовых документов, удовлетворяющих информационным потребностям. Положим, что эти документы были ранее найдены и потому известны. Возьмем большую текстовую коллекцию (H) по исследуемой области с высокой степенью полноты и малой точности относительно пертинентности. Это сделать не очень сложно, причем так что $D_{benefit} \subset H$. Осуществим кластеризацию H посредством латентно-семантического анализа с учетом нечетких коллокаций и получим семейство множеств $H_1, H_2, ..., H_m, m \in N$, причем $H_i \subset H, \forall i = \overline{1,m},$ а $\bigcup_{i=1}^{m} H_i = H$.

Пусть d – исследуемый текстовый документ. Стоит задача определить вероятность $\left(P_{benefit}(d)\right)$ того, что этот документ удовлетворяет информационным интересам субъекта информационного поиска. Благодаря тому, что рассматриваемая коллекция H, а также кластеры, на которые она была разбита $(H_1, H_2, ..., H_m,\ m \in N)$ достаточно

велики, согласно построению, вероятность $P_{benefit}(H_i),\ i=\overline{1,\ m},$ того, что документ удовлетворяет информационным потребностям при условии попадания в некоторый произвольный кластер $H_i, \forall i=\overline{1,m},$ допустимо оценить как частоту таких документов в соответствующем кластере. То есть

$$P_{benefit}(H_i)=\left|D_{benefit}\cap H_i\right|/\left|H_i\right|, \forall i=\overline{1,m}, \tag{4}$$

где $|\cdot|$ – мощность множества.

Обозначим $P_d(H_i),\ i=\overline{1,\ m},$ вероятность того, что документ *d* должен принадлежать кластеру H_i. Напомним, что документ *d* не может быть отнесен к одному из множеств $H_1, H_2, ..., H_m$ посредством процесса кластеризации в силу того, что имеет малый размер. Поэтому задача поиска $P_d(H_i),\ i=\overline{1,m},$ не является тривиальной.

Вместе с тем, если решить данную задачу, то в силу того, что принадлежность к одному из кластеров является полной группой несовместных событий [4] верно, что

$$P_{benefit}(d)=\sum_{i=1}^{m}P_{benefit}(H_i)P_d(H_i). \tag{5}$$

Формулы (4) и (5) позволяют вычислить искомую вероятность при условии того, что известна вероятность $P_d(H_i),\ i=\overline{1,m}$. Последнюю удобно найти посредством теоремы Байеса [4]

$$P_d(H_i)=P(H_i)P_{H_i}(d)\Big/\sum_{k=1}^{m}P(H_k)P_{H_k}(d),\ \forall i=\overline{1,m}, \tag{6}$$

где $P(H_i)$ – априорная вероятность попадания в кластер H_i, которую по аналогии с (4) легко посчитать как

$$P(H_i)=\left|H_i\right|/\left|H\right|, \forall i=\overline{1,m}, \tag{7}$$

а $P_{H_i}(d)$ – вероятность появления документа *d* при условии, что он принадлежит кластеру H_i, которая вычисляется как произведение вероятностей появления его характеристических объектов в кластере H_i.

Данные вероятности в силу большого объема кластеров можно принять равными соответствующим частотам.

Как уже было отмечено ранее, количество коллокаций может быть достаточно велико даже в рамках небольшого текстового документа. Вместе с тем, латентно-семантический анализ позволяет выявить и работать только с теми, которые имеют высокую семантическую значимость.

Процесс извлечения информации из текста в рамках модели (4) – (7) представляет собой поиск непрерывной последовательности предложений в тексте с максимальным значением $P_{benefit}(d)$. Алгоритм тривиален и аналогичен поиску интервала в числовом массиве с максимальной суммой.

Список литературы

1. ГОСТ 7.73–96. Поиск и распространение информации. Термины и определения. Система стандартов по информации, библиотечному и издательскому делу. – Взамен ГОСТ 7.27–80; введ. 01.01.1998 в РФ. – Минск, 2001. – Режим доступа : http://www.docload.ru/ Basesdoc/6/6316/ index.htm.

2. Ягунова, Е. В. Природа коллокаций в русском языке. Опыт автоматического извлечения и классификации на материале новостных текстов / Е. В. Ягунова, Л. М. Пивоварова // Сб. НТИ. – Сер. 2. – № 6. М., 2010. – Режим доступа : http://webground.su/services.php?param= priroda_collac&part=priroda_collac.htm.

3. Поляков, Д. В. Метод формализации нечетких коллокаций термов в текстах на основе лингвистических переменных / Д. В. Поляков, Н. М Митрофанов, А. С. Матвеева. // Прикаспийский журнал. Управление и высокие технологии. – 2015. – № 4(32) – С. 167 – 183.

4. Гмурман, В. Е. Теория вероятностей и математическая статистика / В. Е. Гмурман. – М. : Высшее образование, 2005. – 479 с.

Кафедра «Информационные системы и защита информации» ФГБОУ ВО «ТГТУ»