А. М. Федотов^{1,2}, О. В. Прозоров², О. А. Федотова^{2,3}, А. А. Бапанов⁴

¹ *Институт вычислительных технологий СО РАН
пр. Академика Лаврентьева, 6, Новосибирск, 630090, Россия*

² *Новосибирский государственный университет
ул. Пирогова, 2, Новосибирск, 630090, Россия*

³ *Государственная научно-техническая библиотека СО РАН
ул. Восход, 15, Новосибирск, 630090, Россия*

⁴ *Евразийский национальный университет им. Л. Н. Гумилева
ул. Сатпаева 2, Астана, 010000, Республика Казахстан*

fedotov@nsc.ru

О ПОДХОДЕ К ТЕМАТИЧЕСКОЙ КЛАССИФИКАЦИИ ДОКУМЕНТОВ

Работа посвящена анализу подходов и алгоритмов классификации текстовых документов. Рассматривается подход к тематической классификации документов. Для этого используется специально построенная мера близости документов, учитывающая специфику предметной области. Значения весовых коэффициентов в формуле для вычисления меры близости определяются предполагаемой априорной достоверностью данных соответствующей шкалы.

Ключевые слова: документ, координатное индексирование, мера близости, номинальная шкала.

Введение

В процессе научной и, особенно, образовательной деятельности много времени и сил отнимает работа с литературными источниками: поиск необходимых документов, систематизация документов в соответствии с поставленной задачей. Для удовлетворения информационных потребностей современных пользователей необходима поддержка сложных функций поиска и классификации информации. Наиболее важной является задача систематизации ресурсов, для решения ее необходимо четко определить состав логико-семантических категорий (фасетов) и ключевых терминов (понятий), характеризующих эти категории.

Систематизация ресурсов осуществляется с помощью алгоритмов классификации и кластеризации.

Перед классификацией текстовые документы проходят этап координатного индексирования, после которого основное смысловое содержание текста выражается вектором весов или частот полноточных слов (ключевых терминов). Следующим этапом после индексирования является этап классификации. На этом этапе необходимо определить алгоритм и меру близости, которые будут использоваться для классификации множества текстовых документов. В настоящее время существует множество алгоритмов, используемых для классификации документов: это алгоритмы, использующие разные меры близости, например алгоритм k -ближайших соседей, алгоритм Роккио и др., и алгоритмы, использующие свои понятия о близости документов, например алгоритм Байеса, алгоритм деревьев принятия решений и др.

Федотов А. М., Прозоров О. В., Федотова О. А., Бапанов А. А. О подходе к тематической классификации документов // Вестн. НГУ. Серия: Информационные технологии. 2017. Т. 15, № 1. С. 79–88.

Работа посвящена анализу подходов и алгоритмов классификации текстовых документов.

Постановка задачи

1. Пусть имеется множество документов $D = \{d_n\}_{n=1}^N$, где N – общее количество документов, которое разбито на K множеств классов $C = \{C_k\}_{k=1}^K$, где K – количество классов. Множества C_k , вообще говоря, могут пересекаться.

2. Множество документов D , разбитое на K множеств классов C , может быть подготовлено двумя способами:

- эмпирически;
- аксиоматически.

При эмпирическом способе задание множества классов осуществляется экспертом на основе непосредственного сравнения общих свойств некоторого класса документов. Аксиоматическое задание множества классов осуществляется на основе анализа множества документов D при помощи ранее выработанных условий или правил, которым должен удовлетворять документ. На практике используется комбинация этих способов – над частью документов задается аксиоматическое отнесение к классам, а для документов, загружаемых пользователями, – эмпирическое.

Задачей классификации является отнесение нового документа a к какому-либо классу множества C . Каждый поступающий в систему документ должен автоматически соотноситься с одним или несколькими классами. Множество документов D называют обучающей выборкой.

Координатное индексирование

Индексирование – процесс выражения главного предмета или темы текста какого-либо документа в виде множества ключевых слов или дескрипторов [1]. Применяется для облегчения поиска необходимого текста среди множества других. Различают два основных типа индексирования: классификационное и координатное.

При классификационном индексировании, или классифицировании, тексты в зависимости от их содержания включаются в соответствующий класс (один или несколько), в котором собираются все тексты, имеющие в основном одинаковое смысловое содержание. Каждому такому тексту присваивается индекс этого класса, служащий далее его поисковым образом.

При координатном индексировании основное смысловое содержание текста выражается перечнем однозначных слов, выбираемых либо из самого текста или его заглавия, либо из специального нормативного словаря (тезауруса). В первом случае такие лексические единицы называются ключевыми словами, а во втором – дескрипторами [1]. Каждое ключевое слово или дескриптор обозначает класс, в который потенциально входят все тексты, где в выражения основного смыслового содержания входит это слово. Логическое произведение классов, которые обозначены всеми словами, выражающими в совокупности основное смысловое содержание текста, образует некоторый сложный класс. Построенный таким способом сложный класс обозначается перечнем ключевых слов или дескрипторов, и этот перечень служит поисковым образом данного текста.

Основное преимущество координатного индексирования перед классификационным заключается в том, что координатное индексирование не создает никаких затруднений при поиске текстов по любому, заранее не предусмотренному сочетанию признаков.

Координатное индексирование состоит из нескольких этапов:

- 1) морфологический анализ текстов;
- 2) формирование списка ключевых терминов;
- 3) построение признакового пространства.

Морфологический анализ текстов (лемматизация). При автоматизации процесса координатного индексирования возникает задача идентификации слов текста на естественном языке, приведение их к нормализованному виду и сопоставление их с ключевыми терминами

из тезауруса. Она решается с использованием морфологического анализатора, в котором задаются морфологические признаки слов текста и определяются нормальные формы слов.

Морфологический анализ обеспечивает определение нормальной формы, от которой была образована данная словоформа, и набора параметров, приписанных данной словоформе. Это делается для того, чтобы ориентироваться в дальнейшем только на нормальную форму, а не на все словоформы, использовать параметры, например, для проверки согласования слов [2]. Использование тезауруса позволяет учесть синонимы в ключевых терминах [3].

Формирование списка ключевых терминов. Список ключевых терминов является основным ресурсом для работы алгоритма классификации или кластеризации, помимо самих текстовых документов.

В работе использовалось два способа формирования списка ключевых терминов:

- 1) на основе тезауруса;
- 2) на основе ключевых слов из текстовых документов обучающей выборки D .

При формировании списка ключевых терминов на основе тезауруса берется весь список ключевых терминов тезауруса, а при формировании списка ключевых терминов на основе ключевых слов из текстового документа задается порог встречаемости, начиная с которого слово попадает в список ключевых терминов.

Обозначим множество ключевых терминов, характеризующих тематику документов $T = \{t_i\}_{i=1}^M$, (где M – общее количество ключевых терминов).

Построение признакового пространства. Признаковое пространство определяет данные, которые будут использованы для классификации или кластеризации. Примерами являются частоты встречаемости или веса ключевых терминов.

В случае построения признакового пространства на основе частот встречаемости ключевых терминов выполняется следующее: для каждого ключевого термина подсчитывается частота встречаемости этого термина в каждом отдельном текстовом документе. Эта частота используется для определения степени важности конкретного ключевого термина для текстового документа. Таким образом, для каждого текстового документа формируется вектор с частотами встречаемости ключевых терминов в этом документе. На основе этого вектора формируется некая «точка данных», соответствующая текстовому документу и представляемая в M -мерном пространстве, где M равно количеству ключевых терминов. У каждого текстового документа свой вектор. Именно эти векторы образуют множество объектов, которые и использует на входе алгоритм классификации или кластеризации для разделения текстовых документов.

Кроме частоты встречаемости ключевых терминов, показателем важности ключевого слова для текстового документа является вес. Вес ключевого термина может определяться разными способами. Он может быть задан экспертом или подсчитан на основе частот встречаемости. При подсчете веса на основе частот встречаемости может быть использована технология *TF-IDF*, которая определяет, что вес ключевого термина прямо пропорционален частоте встречаемости этого ключевого термина в текстовом документе, для которого определяется вес, и обратно пропорционален частоте встречаемости этого ключевого термина во всех остальных текстовых документах.

Схема взвешивания TF (term frequency – частота термина). Пусть fr_{ij} – число вхождений термина t_i в документ d_j . Тогда частота термина определяется как

$$TF(t_i, d_j) = \frac{fr_{ij}}{\sum_i fr_{ij}},$$

где $0 \leq i \leq M$, $0 \leq j \leq N$. Отметим, что эта характеристика пропорциональна частоте встречаемости термина.

Схема взвешивания IDF (inverse document frequency – обратная частота термина) – инверсия частоты, с которой термин встречается в документах коллекции. Учет *IDF* уменьшает вес широкоупотребительных слов. Для каждого уникального термина в пределах конкретной коллекции документов существует только одно значение *IDF*:

$$IDF(t_i, D) = \log \frac{N}{|(d_i \subset t_i)|},$$

где N – количество документов в коллекции, $|(d_i \subset t_i)|$ – количество документов, в которых встречается t_i (когда $fr_i \neq 0$), $0 \leq i \leq M$.

Априорный вес некоторого термина пропорционален количеству употребления этого термина в документе и обратно пропорционален частоте употребления термина в других документах коллекции. Кроме того, априорный вес термина зависит от экспертной оценки его значимости.

Учет разделов текста. Документы помимо основного текста имеют некоторое структурное описание, называемое метаданными. К метаданным относятся: заголовок, список авторов, аннотация, авторские ключевые слова, классификационные коды, список литературы и, возможно, другие поля. При классификации эти разделы по другому обрабатываются. В работе используется индексация следующих типов метаданных: заголовок, аннотация, авторские ключевые слова, каждое со своим весовым коэффициентом.

Можно использовать список литературы, выделяя классы, которым принадлежит конкретная ссылка, но у нас для этого слишком мала обучающая выборка (394 документа). Индексация списка литературы ухудшает процедуру классификации.

Мера близости

Для того чтобы сравнивать два объекта, необходимо иметь критерий, на основании которого будет происходить сравнение, т. е. необходимо выбрать меру близости. Как правило, таким критерием является расстояние между объектами.

Мера близости определяет понятие сходства между сравниваемыми объектами. Причем сходство объектов может быть симметричным и несимметричным. Из названий ясно, что, допустим, симметричное сходство определяется тем, что для двух сравниваемых объектов A и B выполняется следующее: если A похож на B , то и B похож на A .

Если же признак, по которому определяется сходство, является отношением, то такое сходство может быть несимметричным. Для полного понимания этого утверждения можно привести пример. На множестве городов – административных центров субъектов Российской Федерации установим в качестве признака сходства близость расстояний между городами. С помощью географической карты нетрудно увидеть, что для Калининграда наиболее близкими (а следовательно, и сходными в заданном понимании) городами будут Псков и Смоленск. Однако для Пскова в качестве «наиболее сходных» городов будут названы отнюдь не Калининград, а Новгород и Санкт-Петербург, для Смоленска – Брянск и Калуга.

Таким образом, сходство определяется мерой близости. Выбор меры близости является важным моментом исследования, от которого решающим образом зависит окончательный вариант разбиения объектов на классы при заданном алгоритме разбиения. В каждой конкретной задаче этот выбор должен производиться по-своему в зависимости от целей исследования, физической и статистической природы вектора анализируемых объектов.

Предлагаемый подход к построению меры близости, используемой для классификации документов, основан на понятии толерантности документов [4].

Количественная характеристика меры близости определяется на множестве документов D следующим образом:

$$m : D \times D \rightarrow [0, 1],$$

причем функция m в случае полного сходства принимает значение 1, в случае полного различия – 0. Рассмотрим два документа d_1 и d_2 .

Пусть $P = \{p_i\}_{i=1}^{\overline{M}}$ – упорядоченный каким-либо образом (например, лексикографически) список ключевых терминов, входящих в оба документа, с учетом повторений (где $\overline{M}$ – общее количество ключевых терминов, с учетом повторений). Вычисление меры близости осуществляется по следующей формуле:

$$m(d_1, d_2) = \sum_{i=1}^{\overline{M}} \alpha_i m_i(d_1, d_2),$$

где i – номер элемента метаданных (ключевого термина), $m_i(d_1, d_2)$ – мера близости по i элементу (иными словами по i шкале), α_i – весовые коэффициенты. Поскольку в описываемой ситуации практически все шкалы – номинальные (состоящие из дискретных текстовых значений), то мера сходства по i -й шкале определяется следующим образом: если значения i -х элементов документов совпадают, то мера близости равна 1, иначе 0.

Весовые коэффициенты должны удовлетворять следующим условиям:

$$\sum_{i=1}^{\overline{M}} \alpha_i = 1, \quad \alpha_i = \alpha_j,$$

если значение термина p_i совпадает с значением термина p_j .

Пусть $T = \{t_k\}_{k=1}^M$ – список уникальных ключевых терминов (список ключевых терминов, формирующих признаков пространство), входящих в оба документа, M_k – число повторений термина t_k . Тогда меру близости можно переписать:

$$m(d_1, d_2) = \sum_{k=1}^M (\alpha_k \cdot M_k) (m_k / M_k),$$

где

α_k – весовой коэффициент соответствующий значению термина t_k ,

m_k – число совпадений термина t_k в документах d_1 и d_2 .

Мы получаем новые весовые коэффициенты $\beta_k = \alpha_k \cdot M_k$, которые характеризуют уже конкретный ключевой термин.

Не трудно видеть, что

$$\sum_{k=1}^M \beta_k = 1.$$

Отметим, что здесь мы автоматически получаем, что весовой коэффициент пропорционален частоте встречаемости термина. Кроме того, при задании меры можно принять во внимание тот факт, что значения весовых коэффициентов β_k определяются предполагаемой априорной достоверностью данных соответствующей шкалы, и в определенных случаях один из коэффициентов может быть увеличен с пропорциональным уменьшением остальных. Например, полное (или даже «почти полное») совпадение значений какого-либо атрибута документа d_1 и документа d_2 может быть более весомо в случае, когда количество значений этого атрибута в документе d_1 достаточно велико (по сравнению со случаем, когда документ d_1 имеет всего одно значение).

Расстояние Евклида. Наиболее распространенная функция расстояния. Представляет собой геометрическое расстояние в многомерном пространстве:

$$p(d_1, d_2) = \sqrt{\sum_{k=1}^M (x_{1k} - x_{2k})^2},$$

где

- 1) x_{1k}, x_{2k} – частота встречаемости (или вес) k -й термина 1-го (2-го) документа;
- 2) $k = 1, 2, \dots, M$;
- 3) M – количество ключевых терминов (размерность пространства).

Широко используется в классификационных алгоритмах в качестве меры близости.

Алгоритмы, использующие меру близости

Пусть имеется множество документов $D = \{d_n\}_{n=1}^N$, где N – общее количество документов, которое разбито на K множеств классов $C = \{C_k\}_{k=1}^K$, где K – количество классов.

Алгоритм k -ближайших соседей. Данный алгоритм является одним из самых простых алгоритмов классификации. Для классификации нового текстового документа необходимо выполнить следующие действия:

- 1) вычислить расстояние до каждого из документов множества D ;
- 2) отобрать k документов, расстояние до которых минимально;
- 3) класс классифицируемого объекта – это класс, который наиболее часто встречается среди k ближайших соседей.

Таким образом, суть алгоритма k -ближайшего соседа: определение отношения документа к преобладающему классу его k соседей. Например, при параметре $k = 1$ алгоритм относит документ к классу самого ближайшего ему документа.

При $k = 1$ качество классификации будет низким, так как классификация зависит только от одного документа из обучающей выборки, а он может иметь неверную метку класса или быть нетипичным. При подборе параметра k обычно полагаются на опыт экспертов или на имеющиеся знания о решаемой задаче. Кроме того, число соседей можно подобрать на обучающем множестве так, чтобы максимизировать качество классификации.

Кроме точности классификации проблемой этого алгоритма является скорость классификации: если в обучающей выборке N документов, подлежит классификации K документов, а размерность пространства M , то количество операций для классификации может быть оценено как $O(K \cdot M \cdot N)$.

Алгоритм Роккио. Данный алгоритм во многом аналогичен алгоритму k -ближайших соседей. Основным отличием является то, что алгоритм Роккио использует центроиды, которые соответствуют классам из обучающей выборки. Для каждого класса определен центроид или представитель класса (представитель класса обычно определяется с помощью эксперта). Центроидом класса C_k называется усредненный вектор частот встречаемости терминов или весов, который вычисляется по формуле

$$\bar{\mu}(C_k) = \frac{1}{|D_{C_k}|} \sum_{d \in D_{C_k}} \bar{v}(d),$$

где D_{C_k} – множество документов класса C_k , $\bar{v}(d)$ – нормализованный вектор документа d .

Правило классификации заключается в определении центроида $\bar{\mu}(C_k)$, к которому точка лежит ближе, чем к другим центроидам, и приписывании этой точки к классу C_k . Другими словами, новые документы классифицируются по ближайшему к ним центроиду.

Классификация Роккио проста в реализации, эффективна по скорости и весьма точна (точность зависит от выбора весовых коэффициентов задании меры близости и представителя класса – насколько точно он описывает класс).

Алгоритм «наивной» классификации. Пусть $T = \{t_m\}_{m=1}^M$ – список уникальных ключевых терминов, характеризующих классы $C = \{C_k\}_{k=1}^K$. На практике ключевые термины выбираются из тезауруса, характеризующего предметную область.

В каждом классе зададим с помощью эксперта веса ключевых терминов β_m^k как веса, используемые по мере близости

$$\mu(d_1, c_k) = \sum_{m=1}^M \beta_m^k \mu_m(d_1, c_k), \quad \sum_{m=1}^M \beta_m^k = 1,$$

где $\mu_m(d_1, c_k)$ – мера сходства между документом и классом по m -му ключевому термину. Поскольку в описываемой ситуации все шкалы – номинальные (состоящие из дискретных текстовых значений), то мера сходства по m -й шкале определяется следующим образом: если

значения m -х элементов совпадают (ключевой термин есть и в документе, и в характеристике класса), то мера близости равна 1, иначе 0.

Алгоритм достаточно прост. Вычисляются меры сходства между новым документом и всеми классами и выбирается класс, с которым сходство максимальное.

На практике документ принадлежит классу C_k , если мера близости превышает определенный порог. Экспериментально установлено, что этот порог $> 0,7$.

Логически этот алгоритм близок к алгоритму дерева принятия решений.

Алгоритмы, не использующие меру близости

Алгоритм «наивной» байесовской классификации. Байесовская «наивная» классификация основывается на следующих предположениях [5]:

1) использование всех ключевых терминов и определение всех зависимостей между ними;

2) наличие двух предположений относительно ключевых терминов:

- все ключевые термины являются одинаково важными;
- все ключевые термины являются статистически независимыми, т. е. вхождение одного ключевого термина в класс не зависит от другого.

Рассмотрим вероятность вхождения ключевого термина $t \in T$ в класс C_k :

$$P(t|C_k) = \frac{H_{tC_k}}{\sum_{t' \in T} H_{t'C_k}}, \quad (1)$$

где H_{tC_k} – частота встречаемости ключевого термина t в документах класса C_k .

Другими словами, числитель описывает, сколько раз ключевое слово t встречается в документах класса C_k (включая повторы), а знаменатель – это суммарное количество встречаемости ключевых терминов во всех документах этого класса.

Определим условную вероятность $P(a|C_k)$ встретить документ a в классе C_k . Алгоритм Байеса представляет документ как набор ключевых терминов. Сделаем предположение, что условные вероятности встречаемости ключевых терминов в документе не зависят друг от друга. Исходя из этого, условную вероятность встречаемости документа можно определить как суммарную вероятность вхождения ключевых терминов документа a в класс C_k .

$$P(a|C_k) = P(t_1|C_k)P(t_2|C_k) \dots P(t_{M(a)}|C_k) = \prod_{t \in T(a)} P(t|C_k), \quad (2)$$

где $T(a)$ – множество ключевых терминов, входящих в документ a .

Однако для этой формулы, если хотя бы один из ключевых терминов, входящих в документ a , не входит в класс C_k , то произведение $\prod_{t \in T(a)} P(t|C_k)$, а следовательно, и $P(a|C_k)$ будут равны нулю. Это приведет к тому, что документ с этим ключевым термином нельзя будет классифицировать, так как он будет иметь нулевую вероятность для этого класса, что неверно, поскольку всегда существует вероятность, что документ принадлежит данному классу. Решением этой проблемы является аддитивное сглаживание (сглаживание Лапласа), идея которого заключается в допущении того, что хотя бы один из ключевых терминов из документа должен присутствовать в классе, т. е. прибавляем единицу к частоте каждого ключевого термина в формуле (1):

$$P(t|C_k) = \frac{H_{tC_k} + 1}{\sum_{t' \in T} H_{t'C_k} + 1} = \frac{H_{tC_k} + 1}{M + \sum_{t' \in T} H_{t'C_k}}.$$

Данный подход дает смещение оценки вероятностей в сторону менее вероятных исходов. Таким образом, ключевые слова, которые не присутствуют в классе, получают пусть маленькую, но все же не нулевую вероятность.

Априорное распределение документов по классам определяется по формуле

$$P(C_k) = \frac{|C_k|}{N},$$

где $|C_k|$ – количество документов из D , которым приписан класс C_k , N – количество документов в D .

Используя формулу полной вероятности, определяем вероятность принадлежности документа a классу C_k :

$$P(C_k|a) = P(a|C_k)P(C_k).$$

Используя метод оценки максимального правдоподобия, рассчитываем вероятности для всех классов и выбираем класс, который обладает максимальной вероятностью:

$$c_{map} = \arg \max_{c_k \in C} [P(a|c_k)P(c_k)]. \quad (3)$$

Подставив выражение (2) в формулу (3), мы получим

$$c_{map} = \arg \max_{c_k \in C} [P(c_k) \prod_{t \in T(a)} P(t|c_k)].$$

Алгоритм деревьев принятия решений. Результат классификации, а также сам ее процесс, наглядно моделируется с помощью алгоритма деревьев принятия решений [5]. Обучающее множество для данного алгоритма представляет собой набор текстовых документов. На основе данного множества строится дерево, которое состоит из следующих составных частей: узлы – это термины документов, листья – метки классов, а ребра – веса терминов.

Пример дерева принятия решений с бинарными весами терминов:

При обучении алгоритма выполняются следующие действия: рассматривают текстовые документы из обучающего множества, проверяют, все ли документы данного множества имеют одинаковую метку класса; если нет, то ищут термин, обладающий наибольшей различительной способностью для разделения этих документов на классы; получают два подмножества документов и строят их поддеревья, повторяя все сначала, пока не получат подмножество документов одного класса, тогда добавляют в соответствующее поддерево лист с меткой этого класса.

Классификация текстового документа происходит следующим образом: документ прогоняется по дереву, выбираются ветви, соответствующие терминам документа. В итоге документу присваивается класс, соответствующий достигнутому листу.

Заключение

В ходе работы были проанализированы алгоритмы классификации, рассмотренные алгоритмы протестированы в сочетании с различными мерами близости (расстояние Евклида и мера, основанная на номинальных шкалах).

Также реализован алгоритм классификации Байеса, который дает похожие результаты.

В результате проведенных экспериментов были сделаны следующие выводы:

1) перед процессом индексирования для повышения качества классификации из текстовых документов необходимо удалить список литературы;

2) перед процессом индексирования для алгоритмов, использующих частоты встречаемости ключевых терминов, из текстовых документов необходимо убрать название, аннотацию и авторские ключевые слова, поскольку они нарушают естественную частоту встречаемости (это ограничение не распространяется на алгоритмы «наивной» классификации и дерева решений);

3) классифицируемые текстовые документы должны быть примерно одинакового размера, иначе качество классификации падает по причине большой разницы в количестве ключевых слов (это ограничение не распространяется на алгоритмы «наивной» классификации и дерева решений);

4) списки ключевых слов классифицируемых текстовых документов должны быть нормированы перед процессом классификации;

5) при классификации алгоритмом Байеса может возникать проблема зануления вероятности вхождения текстового документа в некоторый класс, решением которой является аддитивное сглаживание (сглаживание Лапласа).

Также в процессе экспериментов была выявлена интересная закономерность: алгоритм классификации на основе частотной модели объединяет в одну группу статьи, написанные одним и тем же автором по одной и той же тематике. Это происходит по причине того, что каждый автор имеет уникальный словарный запас, который можно рассматривать как список ключевых слов для определения написанных им текстовых документов.

Тематическая классификация использовалась при формировании списков литературы из электронной библиотеки по учебному курсу «Информатика»¹. Лектором была охарактеризована каждая тема курса – ключевыми терминами из тезауруса по информатике. Проставлены коэффициенты значимости терминов для каждой темы, дальше был применен алгоритм «наивной» классификации для отбора публикаций из библиотеки. Публикации в библиотеке были предварительно проиндексированы с учетом тезауруса. В курсе 8 тем, тема характеризуется 5–12 терминами. Тестирование алгоритма дало хорошие результаты.

Алгоритм деревьев принятия решений на этом материале дал похожие результаты.

Список литературы

1. Михайлов А. И., Черный А. И., Гиляревский Р. С. Основы информатики. 2-е изд. М., 1968.
2. Федотов А. М., Тусупов Д. А., Самбетбаева М. А., Еримбетова А. С., Бакиева А. М., Идрисова А. И. Модель определения нормальной формы слова для казахского языка // Вестн. НГУ. Серия: Информационные технологии. 2015. Т. 13, № 1. С. 107–116. ISSN 1818-7900. EISSN 2410-0420.
3. Федотов А. М., Идрисова И. А., Самбетбаева М. А., Федотова О. А. Использование тезауруса в научно-образовательной информационной системе // Вестн. НГУ. Серия: Информационные технологии. 2015. Т. 13, № 2. С. 86–102. ISSN 1818-7900. EISSN 2410-0420.
4. Федотов А. М., Барахнин В. Б., Жижимов О. Л., Федотова О. А. Модель информационной системы для поддержки научно-педагогической деятельности // Вестн. НГУ. Серия: Информационные технологии. 2014. Т. 12, № 1. С. 89–101. ISSN 1818-7900. EISSN 2410-0420.
5. Большакова Е. И., Клышинский Э. С., Ландэ Д. В., Носков А. А., Пескова О. В., Ягунова Е. В. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика. М.: МИЭМ, 2011. 272 с.

Материал поступил в редколлегию 15.03.2017

¹ <http://fedotov.nsu.ru/info/>

A. M. Fedotov^{1,2}, O. V. Prozorov², O. A. Fedotova³, A. A. Bapanov⁴

¹ *Institute of Computational Technologies SB RAS
6 Acad. Lavrentiev Ave., Novosibirsk, 630090, Russian Federation*

² *Novosibirsk State University
2 Pirogov Str., Novosibirsk, 630090, Russian Federation*

³ *State Public Scientific and Technical Library SB RAS
15 Voskhod Str., Novosibirsk, 630200, Russian Federation*

⁴ *L. N. Gumilyov Eurasian National University
2 Satpaev Str., Astana, 010000, Kazakhstan*

fedotov@nsc.ru

ON THE APPROACH TO THE THEMATIC CLASSIFICATION OF DOCUMENTS

The work is devoted to the analysis of approaches and algorithms for the classification of text documents. The approach to the thematic classification of documents is considered. For this purpose, a specially constructed measure of the proximity of documents is used, taking into account the specifics of the subject area. The values of the weight coefficients in the formula for computing the proximity measure are determined by the assumed a priori reliability of the data of the corresponding scale.

Keywords: document, coordinate indexing, measure of proximity, nominal scale.

References

1. Mihailov A. I., Chernyh A. I., Gilyarevskiy R. S. Fundamentals of Informatics. 2 ed. Moscow, 1968. (in Russ.)
2. Fedotov A. M., Tusupov D. A., Sambetbaeva M. A., Erimbetova A. S., Bakieva A. M., Idrisova I. A. The model for determining the normal form of a word for the Kazakh language. *Vestnik NSU. Series: Information Technologies*, 2015, vol. 13, no. 1, p. 107–116. ISSN 1818-7900. EISSN 2410-0420. (in Russ.)
3. Fedotov A. M., Idrisova I. A., Sambetbaeva M. A., Fedotova O. A. Use of the thesaurus in the scientific and educational information system. *Vestnik NSU. Series: Information Technologies*, 2015, vol. 13, no. 2, p. 86–102. ISSN 1818-7900. EISSN 2410-0420. (in Russ.)
4. Fedotov A. M., Barahnin V. B., Zhizhimov O. L., Fedotova O. A. Model of the information system for supporting scientific and pedagogical activity. *Vestnik NSU. Series: Information Technologies*, 2014, vol. 12, no. 1, p. 89–101. ISSN 1818-7900. EISSN 2410-0420. (in Russ.)
5. Bolshakova E. I., Klyshynskiy E. S., Lande D. V., Noskov A. A., Peskova O. V., Yagunova E. V. Automatic processing of texts in natural language and computer linguistics. Moscow, 2011, 272 p. (in Russ.)

For citation:

Fedotov A. M., Prozorov O. V., Fedotova O. A., Bapanov A. A. On the Approach to the Thematic Classification of Documents. *Vestnik NSU. Series: Information Technologies*, 2017, vol. 15, no. 1, p. 79–88. (in Russ.)