УДК 681.3.01

С.А. Рошка

МЕТОД ИЗВЛЕЧЕНИЯ КЛЮЧЕВЫХ ВЫРАЖЕНИЙ НА ОСНОВЕ ГРАФОВОЙ МОДЕЛИ СТРУКТУРНОГО ПРЕДСТАВЛЕНИЯ ТЕКСТА

1. Введение

Любой документ можно описать с помощью ключевых фраз. Ключевые фразы используются в задачах реферирования и гипертекстового представления текста. Для коллекции документов ключевые фразы могут использоваться для индексирования, классификации, кластеризации, задач обзора документов и поиска [1].

Ключевые выражения дают высокоуровневое описание анализируемого текста. Представление коллекции текстов в виде ключевых выражений позволяет проводить быстрый просмотр документов и их фрагментов на предмет релевантности к тематике, извлекать новые знания из коллекции. Под ключевыми выражениями подразумеваются осмысленные словосочетания из двух или более слов.

Актуальной задачей является разработка алгоритмов автоматизированного извлечения ключевых выражений из текста.

Существуют различные методы решения задачи извлечения ключевых фраз: методы лингвистического анализа [2], статистические методы [3]. В наших исследованиях анализу подвергаются неструктурированные текстовые данные, извлекаемые из WWW и формирующие коллекцию документов. Для обработки такого вида данных целесообразно применять языково-независимые подходы к анализу.

В статье представлен метод извлечения ключевых выражений базирующийся на частотно-статистических методах и контекстном анализе текста. Для применения предложенного метода необходимо использовать графовую модель представления текстовых данных. Графовая модель структурного представления текста предоставляет гибкие возможности для применения методов контекстного и частотно-статистического анализа при исследовании текстовых данных. Представление текста как информационного потока данных позволяет учитывать связанную структуру текста, что обеспечивает для предлагаемого в статье языково-независимого метода извлечения ключевых выражений семантическую значимость результатов, которые получены на сегодняшний день при использовании лингвистических методов обработки текста.

2. Задачи и цели

В настоящее время для задач извлечения ключевых выражений из текста используют методы натуральной языковой обработки текста и технологии машинного обучения. В нашей работе было принято решение использовать методы статистического анализа в силу их относительной простоты, удобства использования и языковой независимости. Методы лингвистического анализа хотя и позволяют точнее анализировать текст, выделяя его структурные особенности, но являются более трудоемкими и сложными в использовании. Связано это, прежде всего, с богатством семантики и морфологии естественных языков. Формальное описание правил естественного языка и их реализация – весьма трудоемкий процесс, требующий привлечения специалистов из области лингвистики. Кроме того, лингвистический анализ предполагает ориентацию на конкретный язык с его конкретными семантическими особенностями, это обусловливает его плохую межъязыковую переносимость.

Все это указывает на целесообразность применения статистических методов для решения задач анализа текста. Однако частотный анализ, используемый в настоящее время при определении тематики документов [4], не позволяет в полной мере учесть внутреннюю структуру текста, т. к. при таком анализе не учитывается связность и последовательность текста. Хотя именно связность текста (речевого высказывания) считается одним из важнейших условий, необходимых для понимания его смысла и содержания. Данное положение является ключевым как в психолингвистике [5, 6], так и нейропсихологии [7, 8].

Опираясь на результаты исследований, авторами был разработан метод извлечения ключевых выражений, основывающийся на модели структурного представления текста и учитывающий его связность [9, 10].

Целью статьи является представление нового метода извлечения ключевых фраз из текста. Метод является частотно-статистическим и опирается на графовую модель представления текста и методы контекстного анализа. Разработанный метод определяет возможности дальнейшего решения задач, таких как: определение тематик в коллекции неструктурированных данных, представление тематик коллекции текстовых данных, определение степени тематической близости текстов.

3. Обзор методов представления текста как информационного потока данных

Общими предпосылками методов представления текста как информационного потока является неопределенность начальных условий для поиска

или анализа информационного контекста. Поэтому методы data mining (интеллектуального анализа данных) будут наиболее подходящими, так как по определению это обнаружение содержательных закономерностей или исключений из данных, возможное без точной фокусировки [11].

В [12] текст рассматривается как последовательности данных. Стартовой точкой являются текстовые данные, конечным продуктом — информация, описывающая явление наличия частот в данных, т. е. фразы или совместно расположенные слова. В предлагаемом методе эта информация представляется как эпизоды или эпизодические правила. Эпизоды и технологии их обработки не рассматривают понимание текста. Вместо этого они разрабатываются для рассмотрения шаблонов, таких как совместное расположение слов или фраз, которые могут использоваться для построения согласованных списков.

Эпизодические правила и эпизоды являются модификацией понятия ассоциативных правил и частотных множеств, примененных к последовательности данных. Последовательность данных, такая как текст, может рассматриваться как последовательность пар (вектор-признак, индекс), где вектор-признак содержит упорядоченный набор признаков, индекс содержит информацию о позиции слова в последовательности. Последовательность представляется в порядке возрастания индекса (соответствуя порядку слов в оригинальном тексте). Признаком может быть: слово — основная форма, флективная форма слова — стем; грамматический признак — часть речи, регистр, номер; знак пунктуации или другой специальный символ или структурный тэг.

В [13] текстовый эпизод рассматривается как пара $\alpha = (V, \leq)$, где V — это коллекция векторов-признаков, и знак $\leq$ — это определенный порядок на V. Дана последовательность текста S, эпизод текста $\alpha = (V, \leq)$ встречается в рамках S, если этот путь удовлетворяет векторам-признакам в V, используя векторы-признаки в S с тем, чтобы частный порядок $\leq$ сохранился. Обычно это означает, что векторы-признаки V могут быть найдены среди S в порядке, удовлетворяющем частному порядку $\leq$.

Для вхождения интересующего эпизода все векторы-признаки эпизода должны полностью совершиться в S. Достаточность покрытия определяется задаваемым пределом, размером окна W, в которых эпизод может встретиться. Поэтому вместо содержания всех эпизодов в S, рассматривается вхождение в подстроку S' от S, где разница определяется векторами-признаками в S' окна W.

Метод обнаружения эпизодов обычно производит большое количество эпизодов и эпизодических правил. Поэтому последующая обработка является существенной для дальнейшей возможности исследований. Основная проблема состоит в том, чтобы определить, какой эпизод является осмысленным. Последующая обработка включает сокращение, группировку и упорядочивание результатов. Практичность результатов может быть усилена с применением знаний на правилах и эпизодах одного документа и сравнением их с подобной информацией в коллекции документов.

Ассоциативные правила, такие как в [14], не рассматривают полный текст как входную информацию, они используют только ключевые слова, приписываемые к этому документу. В [15–18] скомбинированы и развиты эти методы. С одной стороны, для извлечения подходящих фраз из документа для будущей обработки и обнаружения знаний и, с другой — для поиска взаимного расположения между этими фразами. Технология для поиска комплекса текстовых фраз из полного текста позволяет по желанию менять интервал между словами и порядок слов в фразах. Метод не разграничивает по типам фраз, рассматриваются все слова и словоформы. Перед извлечением фраз документ подвергается предварительной обработке, при которой удаляется множество общей и несодержательной информации. Для поиска совместно расположенных фраз создается набор последовательностей слов для каждого единственного документа. Набор фраз может использоваться для описания документа в коллекции. Фраза определяется путем извлечения последовательности, удовлетворяющей частотному порогу δ. Последовательность является частой, если она присутствует не менее чем в δ документах, где δ— это частотный порог.

Снижение количества частотных наборов достигается на этапе предварительного морфологического анализа документов.

В [19] устойчивые ключевые выражения определяются как повторяющиеся упорядоченные последовательности термов [20], присутствующих в документах. Предполагается, что при хорошем стиле изложения используются синонимы и местоимения и нет повторений. Определяются абстрактные понятия – единичные объекты или группы связанных объектов, которые когнитивно отличаются от других абстрактных понятий. Для того чтобы быть подходящим для кластеризации, устойчивая последовательность или единичный терм должны:

1) присутствовать в коллекции не менее определенного количества раз (ограничитель частоты термов);

2) не пересекать границы предложения;

3) быть завершенной фразой;

4) не начинаться и не заканчиваться стоп-словами.

4. Модель структурного представления и метод тематического анализа текста

Всю совокупность представленных на сегодняшний день методов анализа текста, относительно задачи анализа его содержания, можно разделить на две большие группы:

- лингвистический анализ;
- статистический анализ.

Первый ориентирован на извлечение смысла текста по его семантической структуре. Второй – по частотному распределению слов в тексте.

В [7] авторами было принято решение использовать методы статистического анализа в силу их относительной простоты, удобства использования и языковой независимости.

4.1. Графовая модель структурного представления текста произвольного содержания. Суть предлагаемого подхода заключается в моделировании структуры текста информационным потоком и формировании этим потоком ориентированного мультиграфа, вершинами которого являются слова, а ребрами – связи между словами в тексте. Этот мультиграф является информационной структурой текста.

Мультиграф – это граф, который может содержать множество ребер, соединяющих одну и туже пару вершин.

Информационный поток – это детерминированный поток событий, принадлежащих некоторому конечному множеству. Временной интервал между событиями нас не интересует, интересует только последовательность событий. В данном случае события – это слова, а конечное множество – это множество всех слов, присутствующих в анализируемом тексте. Информационный поток эквивалентен временному ряду номинальных (категориальных) величин.

Под информационной структурой понимается совокупность всех событий и связей между ними. Для информационной структуры текста связи между событиями – это словосочетания.

Информационный поток, по сути, моделирует динамику некоторого процесса, в данном случае текста, а информационная структура является статическим представлением информационного потока.

Переход к модели структурного представления текста осуществляется следующим образом:

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

Если последовательно брать слова из текста, начиная с самого первого и заканчивая последним, то это как раз и будет информационный поток F.

При этом набор всех слов в тексте можно выделить в конечное множество уникальных информационных элементов:

$$I=\{i_1,i_2,\ldots,i_n\},$$

где: i – информационный элемент соответствующий определенному слову из текста.

$$F=(i_1,i_2,\ldots,i_n).$$

Информационный поток также может быть представлен в виде набора связей:

$$F=(r_1,r_2,\ldots,r_{n-1}),$$

где: $r_i=(i_i,i_{i+1})$ – связь между двумя информационными элементами, последовательно идущими в информационном потоке.

Порядок чередования информационных элементов зависит от их последовательности в тексте. Информационные элементы в потоке могут повторяться. Обязательное условие – однозначное соответствие информационного элемента слову из текста. Одинаковые слова в тексте соответствуют одному и тому же информационному элементу.

Пример.

Фрагмент текста: «Дао, которое может быть выражено словами не есть постоянное Дао. Имя, которое может быть названо, не есть постоянное имя».

Из данного набора слов выделяем множество уникальных информационных элементов (различия в регистре и знаки препинания не учитываются): $I=\{i_1,i_2,i_3,i_4,i_5,i_6,i_7,i_8,i_9,i_{10},i_{11}\}$, где:

i_1 = быть,
i_2 = выражено,
i_3 = дао,
i_4 = есть,
i_5 = имя,
i_6 = которое,
i_7 = может,
i_8 = названо,
i_9 = не,
i_{10} = постоянное,
i_{11} = словами.

2. Поток формирует структуру.

Если учесть, что слова в тексте повторяются, то, соответственно, можно допустить, что информационный поток будет многократно проходить через одни и те же информационные элементы, формируя таким образом связанную информационную структуру текста.

Введем дополнительные обозначения и определим некоторые важные характеристики информационной структуры:

$n(I)=|I|$ — количество информационных элементов множества I (количество уникальных слов в тексте).

$n(F)=|F|$ — количество информационных элементов набора F (общее количество слов в тексте).

$M(I, R)$ — информационная структура (ориентированный мультиграф). Является совокупностью

I — множества информационных элементов (вершин графа) и R — набора связей между этими элементами (ребер графа):

$$M(I, R) \equiv F,$$

R — набор связей между парами информационных элементов; может содержать повторяющиеся связи в случае многократного прохождения информационного потока F через одни и те же пары элементов:

$$R = (r_1, r_2, \ldots, r_{n-1}),$$

где: $r_i = (i_i, i_{i+1})$ – связь между двумя информационными элементами; обозначает последовательность информационных элементов i_i, i_{i+1} в потоке F.

Для каждого информационного элемента из множества I, входящего в структуру $M(I, R)$, существует набор пар связей, где: r_i, r_{i+1} – входная связь, r_{i+1}, r_{j+1} – выходная связь, n – число пар связей.

$$\forall\, i \in I\ \exists\ R(i) = ((r_i, r_{i+1})_1, \ldots, (r_j, r_{j+1})_n).$$

Входная означает, что данная связь предшествует выходной в наборе связей, описывающих поток, проходящий через данный информационный элемент. Если проиндексировать связи в наборе, описывающие поток, то индекс входной связи будет на единицу меньше выходной.

$n(R(i))$ — количество пар связей в наборе $R(i)$.

$n(R(i))$ характеризует число связей данного информационного элемента с другими информационными элементами в структуре $M(I, R)$. $n(R(i))$ равно числу повторений слова в тексте.

Информационный поток относительно некоторого информационного элемента i можно описать как

$$F(i, e, [r^-, r^+]),$$

где: e – вхождение информационного элемента i_k в поток F, его порядковый номер в потоке; $[r^-, r^+]$ – окрестность, для которой определяется поток:

$$F(i, e, [r^-, r^+]) = (i - r^-, \ldots, i-2, i-1, i, i+1, i+2, \ldots, i+r^+),$$

где: $i \pm n$ — обозначает индексацию некоторого информационного элемента в наборе F относительно информационного элемента i; $i+1$ – обозначает информационный элемент, следующий сразу за i в информационном потоке F; $i-1$ – обозначает информационный элемент, предшествующий i в информационном потоке F.

Обозначим множество всех информационных потоков относительно информационного элемента i для всех его вхождений в поток F:

$$D(F, i, [r^-, r^+]) = \bigcup_{e=0}^{d(i)} F_e(i, e, [r^-, r^+]).$$

На базе представленной модели опишем метод из задачи тематической классификации текстовой информации – извлечение ключевых выражений из коллекции неструктурированных текстовых данных.

4.2. Извлечение ключевых выражений из коллекции текстовых данных с использованием метода частотно-контекстной классификации. Предлагаемый подход к тематической классификации текстовой информации основывается на гипотезе о том, что словарный запас и частота использования слов зависят от темы текста [21]. В настоящее время данная гипотеза активно и успешно используется в тематико-ориентированных методах поиска [22].

Тематическая классификация предполагает выделение множества ключевых выражений, определяющих тематику текста. При этом каждому из них приписывается вес, определяющий значимость данного выражения в тематике, т. е. какие-то ключевые выражения играют большую роль в определении тематики, какие-то меньшую, но именно такая совокупность ключевых выражений, с такой значимостью каждого из них в тематике и определяет тематическую направленность.

Ключевые выражения определяются по количеству их вхождений в текст, а именно – частота ключевых выражений в тексте определяется пороговой величиной ∂. В рамках рассматриваемой модели структурного представления текста это будет означать, что через данные ключевые выражения чаще проходит информационный поток, и информационные элементы, соответствующие этим словам, имеют большее количество связей с другими информационными элементами.

Общая последовательность метода будет выглядеть следующим образом:

1. Моделирование текста и формирование информационной структуры $M(I, R)$.

2. Выделение множества всех информационных элементов, ранжированных по их степени d(i) (числу повторений в тексте). Элемент с $d(M(I, R))_{\max}$ будет первым, и далее по убыванию.

3. Выделение множества ключевых элементов.

Нашей задачей является выделение максимально повторяющихся последовательностей информационных элементов из множества всех информационных элементов.

Введем пороговую величину ∂ — минимальное количество повторений выделенных последовательностей, при которой последовательность является значимой. Это множество последовательностей будет формировать набор ключевых элементов.

Берем n первых элементов, n определяется на основе пороговой величины ∂:

$$S_k = \{ k_1 I_1, k_2 I_2, \ldots, k_n I_n \},$$

$$I_1 = \{ i_h, \ldots, i_m \},$$

$$I_i = \{ i_o, \ldots, i_l \},$$

где I_i – набор информационных элементов информационного потока F, $n(I_i)$ – количество элементов в наборе I_i,

$$1 \le n(I_i) \le n(F).$$

Коэффициенты k_1, k_2,..., k_n соответствуют степеням информационных множеств

$$I_i = d(I_i),$$

где: $d(I_i)$ – количество повторений множества информационных элементов I_i в потоке F; S_{k_i} – набор информационных элементов I, входящий в набор $\bigcup I_l$, $2 \le l \le n(F)$.

Одним из ключевых моментов является тот факт, что через информационный элемент поток может проходить множество раз. Тем более это справедливо для первичного множества ключевых элементов I, так как именно они выбраны из всего текста на том основании, что у них больше связей с другими информационными элементами.

Окрестность информационного элемента I, а это множество информационных элементов, входящих в $D(F, i, [r^-, r^+])$, в этом случае является контекстом. Анализ окрестности информационного элемента для выделения контекста данного элемента будем называть контекстным анализом.

Формирование множества ключевых выражений – извлечение максимальных повторяющихся последовательностей из информационного потока F выглядит следующим образом:

1. Для формирования окрестности r – зададим информационные потоки, которые будут формироваться в результате использования предлагаемых методов в [21]. Они позволяют упростить методы работы с информационными потоками и дают величины для значения параметров математической модели обработки текста как информационного потока.

При заключении текстовых данных коллекции в предлагаемую нами структуру данных мы получаем множество m информационных потоков F_c — это понятия, на которые мы разделяем текст коллекции по модели данных.

Весь текстовый поток данных коллекции обозначим как F:

$$F = F_c = \bigcup_{i=1}^{m} F_{c_i}.$$

Далее для извлечения ключевых выражений каждое понятие разделяется на подпонятия – множества F_{sc}, по всей коллекции у нас получается l подпонятий, которые в информационном потоке F будут являться информационными элементами для F_c.

$$F_{sc} = \bigcup_{i=1}^{l} F_{sc_i},$$

$$\forall\ F_{c_i}\ \exists \bigcup_{j=1}^{b} F_{sc_j} = \{ F_{sc_1}, F_{sc_2}, \ldots, F_{sc_b} \}.$$

Выделяем уникальные термы. Термы являются информационными элементами для всего потока F коллекции и для F_c, F_{sc}.

Обозначим это множество как I.

$n(I) = |I|$ — количество уникальных элементов в F,

$$I = \{i_1, i_2, \ldots, i_n\}.$$

Мы можем записать все информационные потоки как множества информационных элементов в виде:

$$F_{sc_s} = \{ i_l, \ldots, i_j \},$$

$$F_{c_i} = \{\{ i_i, \ldots, i_j \}, \ldots, \{ i_s, \ldots, i_l \}\},$$

пусть $I_j = \{ i_k, \ldots, i_l \}$, тогда $F_{sc_i} = I_j$,

$$F_{sc} = \bigcup_{i=1}^{l} I_i,$$

$$F_{c_i} = \bigcup_{i=1}^{s} F_{sc_i} = \bigcup_{i=1}^{s} \bigcup_{j=1}^{l} I_j,$$

где: s — количество подпонятий F_{sc} в i-м понятии; l — количество информационных элементов в каждом j-м подпонятии.

$$F_c = \bigcup_{i=1}^{n} F_{c_i},$$

где: n – количество понятий в потоке F.

$$F_c = \bigcup_{i=1}^{n} F_{c_i} = \bigcup_{i=1}^{n} \bigcup_{j=1}^{s} \bigcup_{k=1}^{l} I_k.$$

Выделяем в набор $A(i)$ множество всех потоков, проходящих через каждый информационный элемент $i \in F$ в некоторой окрестности, заданной r.

Для этого, согласно нашей модели данных, так как:

$$F_{sc} = \bigcup_{j=1}^{s} F_{sc_j},$$

то для каждого информационного потока F_{sc_j} запишем множество $A(i)$ всех потоков, проходящих через информационный элемент i:

$$A(i) = D(F_{sc_j}, i, [r^-, r^+]).$$

Окрестность будет переменной величиной в зависимости от $d(F_{sc_j})$. Параметры r^- и r^+ будут различными.

Объединим все наборы $A(i)$, для каждого $i \in F$ в один общий набор $A(F_{sc})$:

$$A(F_{sc}) = \bigcup_{k=1}^{n} A(i_k) \ , k \in F \ , \ 1 \le k \le n(F),$$

обозначим его как $A = A(F_{sc})$.

В результате мы получили общий набор A, включающий в себя все потоки, проходящие через информационные элементы множества F.

Теперь из набора информационных элементов A выделяем множества, которые будут являться ключевыми выражениями S_{k_i}, $I = 1...t$, где t – количество полученных множеств. Общий набор обозначим как P_k, $A \to P_k$.

При этом будем учитывать количество повторяющихся информационных элементов и для каждого элемента S_k запишем число их повторений в наборе A:

$$A: P_k = \{k_1 P_{k_1}, k_2 P_{k_2}, ..., k_n P_{k_n}\},$$

где коэффициенты $k_1, k_2, ..., k_n$ — это число повторений этих информационных элементов в наборе A.

Все элементы множества F_{sc_j} присутствуют в некоторой окрестности r^- и r^+ элементов множества F_c, и каждый информационный элемент $i \in F$ определяет центр некоторой окрестности в информационной структуре $M(I, R)$. Окрестность задается по информационным потокам, проходящим через i.

Число повторений затем используется для определения весов этих слов в тематике.

Целью является извлечение повторяющихся последовательностей информационных элементов на декартовом произведении множеств $F_{sc} \times F_{sc}$ с учетом частотного порога ∂. Полученные потоки информационных элементов p_i будут представлять собой множество:

$$p_i = \{ i_s, ..., i_l \} \ s,...,l \in I.$$

Важен лексикографический порядок следования информационных элементов.

Множество ключевых выражений можно представить в виде:

$$P_k = \{ \lambda_1 p_1, \lambda_2 p_2, ..., \lambda_k p_k \},$$

где: p_i — ключевое i-е выражение, то есть извлеченная повторяющаяся последовательность информационных элементов; λ_i — коэффициент мощности ключевого выражения, то есть количество найденных повторений i-ой последовательности информационных элементов.

5. Заключение

В статье мы провели обзор существующих моделей представления текстовых данных для последующего анализа с целью определения тематик текста, которые на сегодняшний день применяются в технологиях text mining и KDT.

Научная новизна: в данной статье описана графовая модель структурного представления текста произвольного содержания и на ее основе предложен новый метод извлечения ключевых выражений из коллекции неструктурированных текстовых данных.

Практическая значимость определяется возможностью дальнейшего применения модели при решении задач определения тематик в коллекции текстовых данных, сравнении текстов на тематическую близость.

В сравнении с аналогами получили, что частотный анализ, используемый в настоящее время для извлечения ключевых выражений, не позволяет в полной мере получить результаты, близкие к показателям при использовании лингвистических методов обработки текстовых данных, но необходимым условием работы с данными является именно использование языково-независимых алгоритмов. Многими исследователями при частотном анализе не учитывается связность и последовательность текста. Хотя именно связность текста считается одним из важнейших условий, необходимых для понимания его смыла и содержания. Предлагаемый метод извлечения ключевых выражений приближает результаты обработки текста к результатам, получаемым с помощью лингвистических методов и открывает широкие возможности для дальнейших исследований в области анализа неструктурированных данных.

Список литературы: 1. *D'Avanzo E., Margini B., Lavelli A., Zanoli R.* Using Keyphrases as Features for Text Categorization. ITC-irst Technical report, 2003, Ref. No.: T03-11-01, 12 pp. 2. *Семенова С.Ю.* Поиск параметрической информации в тексте: алгоритмический и лексикографический аспекты // Труды Междунар. семинара «Диалог'96 по компьютерной лингвистике и приложениям». М., 1996. С. 227-230. 3. *Church K.W., Gale W., Hanks P., Hindle D.* Using statistics in lexical analysis. In Uri Zernik (ed.), Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon, 1991, pp. 115-164. 4. *Singhal A., Mitra M., and Buckley C.* Learning routing queries in a query zone. In Proc. of the SIGIR'97, July 1997, pp. 25-32. 5. *Белянин В.П.* Введение в психолингвистику. Изд. 2-е, испр. и доп., М.: ЧеРо, 2000. 128 с. 6. *Фрумкина Р.М.* Психолингвистика: Учебник для студ. высш. учебн. заведений. М.: Изд. центр "Академия", 2001. 320 с. 7. *Адарюков В.И.* Исследование и разработка машинно-ориентированного метода инфологического моделирования информационно-поисковых систем фактографического типа: Дисс. ... канд. техн. наук: 05.13.06. ЛЭТИ им. В.И. Ульянова (Ленина). СПб., 1988. 256 с. 8. *Лурия А.Р.* Основы нейропсихологии. М.: МГУ, 1973. 374 с. 9. *Чугреев В.Л., Яков-*

лев С.А. Выделение критериев поиска текста на основе подобия значимых документов // Вузовская наука — региону: Материалы 1-й Общерос. научн.-техн. конф. Вологда: ВоГТУ, 2003. С. 200-202. 10. *Чугреев В.Л., Яковлев С.А.* Анализ структуры текста и прогнозирование нечисловых величин // Там же. С. 202-204. 11. *Helene Ahonen, Oskari Heinonen, Mika Klemettinen, A. Inkeri Verkamo.* Mining in the phrasal frontier. In Jan Komorovski and Jan Zytkow, editors, Proceedings of the First European Symposium on Principles on Data Mining and Knowledge Discovery (PKDD'97), number 1263 in Lecture Notes in Artificial Intelligence, Trondheim, Norway, Jine 1997, pp. 343-350. 12. *H. Mannila, H. Toivonen, and A. I. Verkamo.* Discovering frequent episodes in sequences. In Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD'95). Montreal, Canada, Aug 1995, pp. 210-215. 13. *Agrawal. R., Mannila H., Srikant R., Toivonen H., and Verkamo A.I.* Fast discovery in association rules. Advances in Knowledge Discovery and Data Mining / In Fayyad U.M., Piatetsky-Shapiro G., Smyth P., and Uthurusamy R., editors, AAAI Press, Menlo Park, California, USA 1996, pp. 307-328. 14. *R. Feldman, W. Kloesgen, A. Zilberstein.* Document Explorer: Discoverin knowledge in document collections. Proceedings of Tenth International Symposium on Methodologies for Intelligent Systems (ISMIS'97), number 1325 in Lecture Notes in Artificial Intelligence. Charlotte, North California, USA, October 1997, pp. 137-146. 15. *Frank E., Paynter G.W., Witten I.H., Gutwin C. and Nevill-Manning C.* Domain-specific keyphrase extraction. Int Joint Conference on Artificial Intelligence, Stockholm, Sweden, 1999, pp. 668–673. 16. Ahonen, H., Heinonen, O., Klemettinen, M., and Verkamo. Finding co-occurring text phrases by combining sequence and frequent set discovery. In Proceedings of IJCAI'99 Workshop on Text Mining: Foundations, Techniques and Applications, 1999, pp. 1-9. 17. *H. Mannila, H. Toivonen, and A. I. Vercamo.* Discovering frequent episode in sequences. In Proc. of the 1st International Conference on Knowledge Discovery and Data Mining, AAAI Press, Aug. 1995, pp. 210—215. 18. *Zaki M. J. SPADE:* An Efficient Algorithm for Mining Frequent Sequences, Machine Learning Journal, Vol. 42 Nos. 1/2, Jan/Feb 2001, pp. 31-60. 19. *Chidanand Apte, Fred Damerau, Sholom M. Weiss,* Towards Language Independent Automated Learning of Text Categorisation Models. Research and Development in Information Retrieval, 1994, pp. 23-30. 20. *Lovins J.* Development of a stemming algorithm. Mechanical translation and computational linguistics, Vol. 11, 1968, pp. 22–31. 21. *Некрестьянов И.С.* Тематико-ориентированные методы информационного поиска: Дисс. ... канд. техн. наук: 05.13.11. СПб. гос. ун-т. СПб., 2000. 80 с. 22. *Salton G., Allan J., and Singhal A.* Automatic text decomposition and structuring. Information Processing & Management, 32(2), 1996, pp. 127-138.

Поступила в редколлегию 23.05.2005