Опыт создания кластеров документов на основе метода определения их тематического подобия

© В.Н. Захаров © Александр А. Хорошилов © Алексей А. Хорошилов
ИПИ РАН, ЦИТиС,
Москва
vzakharov@ipiran.ru khoroshilov@mail.ru a.a.horoshilov@mail.ru

Аннотация

В работе описывается опыт применения методов построения формализованного смыслового описания и оценки подобия тематического содержания текстов. Применяемые в исследовании методы базируются на использовании процедур семантико-синтаксического и концептуального анализа, обеспечивающих выявление понятийного состава текста и назначения наименованиям понятий характеристик, соответствующих их семантической роли и значимости в тексте. Для выполнения данной работы был создан комплекс программных средств, который был опробован на англоязычных текстах.

1 Задача установления тематической близости документов

1.1 Введение

В настоящее время в различных фондах накоплены огромные массивы текстовых документов по широкому спектру тематических областей. Для решения различных задач дальнейшего эффективного использования этих документов необходима их предварительная автоматическая обработка, позволяющая свести к минимуму трудозатраты обслуживающего персонала. Одной из ключевых задач обработки текстовой информации является проблема установления смысловой связи между различными документами. Существуют различные методы решения данной задачи, позволяющие с той или иной степенью эффективности ее решить. В данной статье описывается применение одного из методов оценки подобия тематического содержания текстов. Данный метод может быть применён при сравнении документов на различных языках. Ранее уже было описано его применение для русскоязычных текстов [9]. В этой статье мы описываем эксперимент, проводимый на массиве англоязычных документов.

Задача данного эксперимента состояла в проверке работоспособности метода на массиве документов, предоставленных заказчиком. Необходимо было создать кластеры документов близких к документам-образцам. Авторам статьи была передана подборка англоязычных текстов журнала «Science», включающая 1584 документа по широкому спектру тематических областей. В связи с тем, что работа выполнялась за счет внутренних ресурсов и в ограниченные сроки, было принято решение не проводить предварительную подготовку полученных текстовых слоев к проведению эксперимента, несмотря на некоторые недостатки распознавания предоставленных документов. Основные ошибки заключались в разбиении слов на фрагменты и ошибки в словах из-за неверного распознавания. Также при распознавании появлялись дополнительные символы и группы символов не имеющие смысловой нагрузки. Процедура распознавания проводилась на стороне заказчика.

1.2 Существующие методы и средства для создания кластеров документов

Проблеме кластеризации текстов посвящено огромное количество исследований и разработано довольно много методов, которые позволяют решить данную задачу с разной степенью эффективности. Многие авторы пытаются обобщить эти исследования, результатом чего становятся работы, посвященные описанию и сравнению существующих методов. В работах [1–5] описаны такие методы как, например:

- LSA/LSI – Latent Semantic Analysis/Indexing. Путем факторного анализа множества документов выявляются латентные (скрытые) факторы, которые в дальнейшем являются основой для образования кластеров документов;
- STC – Suffix Tree Clustering. Кластеры образуются в узлах специального вида дерева – суффиксного дерева, которое строится из слов и фраз входных документов;

Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13–16 октября 2014 г.

- Single Link, Complete Link, Group Average – эти методы разбивают множество документов на кластеры, расположенные в древовидной структуре – dendrogramm, получаемой с помощью иерархической аггломеративной кластеризации;
- Scatter/Gather. Представляется как итеративный процесс, сначала разбивающий (scatter) множество документов на группы и представлении затем этих групп пользователю (gather) для дальнейшего анализа. Далее процесс повторяется снова над конкретными группами.
- K-means. Относится к не-иерархическим алгоритмам. Кластеры представлены в виде центроидов, являющихся «центром массы» всех документов, входящих в кластер.
- CI – Concept Indexing. Разбивает множество документов методом рекурсивной бисекции, т.е. разделяя множество документов на две части на каждом шаге рекурсии. Метод может использовать информацию, полученную на этапе обучения.
- SOM – Self-Organizing Maps. Производит классификацию документов с использованием самонастраивающейся нейронной сети.

Стоит заметить, что все методы обладают своими достоинствами и недостатками, которые описаны в работах [1–2, 4], но эффективность работы всех методов определяется точностью определения слов и словосочетаний, характеризующих документ. Именно повышение точности определения значимой лексики документа является одной из приоритетных задач наших исследований.

2 Используемые средства обработки текстовой информации

Для выполнения эксперимента, который описан ниже, было использовано программное обеспечение – лингвистический процессор МетаФраз, которое позволяет решать задачи автоматической обработки текстовой информации. Его основной задачей является структурирование и формализация смыслового содержания текстов, выявление понятийного состава предметной области, установление парадигматических, синтагматических и ассоциативных связей между наименованиями понятий и установление их контекстного окружения. Центральными процедурами, использующимися в лингвистическом процессоре, являются процедуры семантико-синтаксического и концептуального анализа текстов [8, 10, 11, 13, 14]. Для установления смысловых связей между текстовыми документами используется метод, описанный в работе [9]. Он был адаптирован для задач обработки англоязычных текстов.

3 Технология составления частотных словарей по корпусу текстов

При решении задач автоматизированного составления словарей важно выявить понятийный состав предметной области для его последующей обработки и включения в состав создаваемого концептуального словаря. При этом, как показывают исследования, любой репрезентативный тематический корпус текстов является по своему лексическому составу политематическим (т.е. в нем присутствует лексика широкого спектра тематических областей), и предметные области отличаются друг от друга не их лексическим составом, а распределением частот появления в них различных наименований понятий. Поэтому технология автоматического составления частотных словарей имеет важное значение и для задачи составления терминологических словарей.

Эту технологию можно представить следующим образом. Предварительно составленный корпус текстов подвергается обработке процедурой семантико-синтаксического и концептуального анализа текстов, в результате чего из текстов выделяются отдельные слова и словосочетания различной длины. Далее по массиву выделенных из текстов наименований понятий составляется частотный словарь. После этого полученный словарь обрабатывается процедурой орфографического и синтаксического контроля, в результате чего из этого словаря исключаются некорректные слова и словосочетания. И, наконец, частотная часть словаря подвергается лингвистической обработке, в результате которой из словаря исключается малоинформативная и некорректная лексика.

Автоматизированное составление словарей наименований понятий можно выполнить по следующей технологической схеме:

1. Формально-логический контроль исходных текстов с целью обнаружения и исправления орфографических и синтаксических ошибок в исходных текстах;
2. Членение исходного текста на отдельные слова (по пробелам и разделительным знакам между ними);
3. Морфологический анализ слов корпуса текстов;
4. Членение корпуса текстов на предложения;
5. Семантико-синтаксического анализ текстов;
6. Приближенный концептуальный анализ текстов;
7. Выделение наименований понятий;
8. Автоматическое приведение наименований понятий к их канонической форме;
9. Формирование частотного словаря наименований понятий;
10. Лингвистический анализ частотного словаря наименований понятий (Исключение ошибочной и малоинформативной лексики);
11. Формирование машинного представления концептуального словаря.

4 Эксперимент по установлению тематической близости массива документов

4.1 Предварительный анализ массива текстов

Для проведения исследований был составлен массив англоязычных текстов журнала «Science» (10 годовых выпусков). Количество текстов в массиве составляло 1835 документов. Их совокупный объем (в символах) составил 33 Мб. Результаты проведенного нами предварительного автоматического анализа этого массива текстов приведены в таблице 1.

Таблица 1

Результаты предварительного анализа массива текстов

Количество текстов	1835
Совокупный объем (в символах)	33 млн
Количество разных словосочетаний	17 млн
Количество разных слов	887 тыс.

Предварительный анализ качества текстов случайной выборки документов показал, что в обрабатываемых документах процент неправильно распознанных слов для различных страниц колебался от 5 до 24%. Данный факт привел к резкому возрастанию в массиве текстов числа ошибочных слов и разрушенных синтаксических конструкций словосочетаний и предложений. Как видно из таблицы 1, значительно возросло общее число слов (по нашим оценкам в аналогичных объемах текстов их число не должно превышать 100-120 тыс. слов), а общее число словосочетаний (по нашим оценкам их число не должно превышать 3-4 млн словосочетаний). Таким образом, представленные для проведения анализа тексты изначально содержали значительную долю ошибочных текстовых ситуаций. Это существенное допущение, которое исказит количественные характеристики анализа текстов.

Ввиду того, что в состав массива текстов, предоставленного для выполнения эксперимента, включены тексты широкого спектра тематических областей и он достаточно большого объема, было решено использовать его в качества корпуса текстов для составления словаря наименований понятий.

В соответствие с приведенной выше технологией был составлен частотный словарь словосочетаний. При этом в силу ограничений по трудозатратам из технологического процесса были исключены п. 1 и 10, а отдельные слова вследствие их многозначности и большого числа искажений в них было решено проигнорировать. В таблице 2 приведен фрагмент частотного словаря наименований понятий. В нем для каждой записи слева указана частота встречаемости наименования понятия в корпусе текстов, наименование понятия, представленное в виде пословно нормализованных слов, и одна из текстовых форм наименования понятия.

Таблица 2

Фрагмент частотного словаря словосочетаний

*00001018 health service * health services*
*00001017 gene transfer * gene transfer*
*00001017 natural resource * natural resources*
*00001015 risk assessment * risk assessment*
*00001013 fusion protein * fusion proteins*
*00001009 blood pressure * blood pressure*
*00001009 guinea pig * guinea pigs*
*00001006 aids patient * AIDS patients*
*00000996 somatic cell * somatic cell*
*00000994 cell body * cell body*
*00000994 muscle cell * muscle cells*
*00000989 national security * national security*
*00000985 growth hormone * growth hormone*
*00000970 protein structure * protein structure*
*00000958 communication skill * communications skills*
*00000955 nuclear weapon * nuclear weapons*
*00000944 mental health * mental health*
*00000943 natural science * natural sciences*
*00000942 human chromosome * human chromosomes*
*00000935 rat brain * rat brains*
*00000935 salary range * salary range*
*00000925 clinical research * clinical research*
*00000923 skeletal muscle * skeletal muscle*
*00000919 research grant * research grants*
*00000916 cell proliferation * cell proliferation*
*00000915 serum albumin * serum albumin*
*00000914 dna polymerase * DNA polymerase*
*00000905 membrane potential * membrane potentials*
*00000901 electron density * electron density*
*00000901 molecular size * molecular size*
*00000900 hela cell * HeLa cells*
*00000886 synthetic peptide * synthetic peptides*
*00000869 low temperature * low temperatures*
*00000868 high energy * high energy*

На основе анализа результатов обработки частотного словаря было принято решение использовать в качестве исходных данных для составления эталонного концептуального словаря только часть частотного словаря с частотами «три и более». Это решение основывалось на следующем допущении: словосочетания фрагмента словаря с частотой равной 1 встречаются однократно и, следовательно, не могут служить инструментом классификации текстов и, кроме того, в них содержится максимальный процент ошибок. Аналогичное допущение в значительной степени может относиться и к части словаря с частотой равной двум. Что касается частоты «три и более», то здесь повторяемость словосочетаний служит некоторой гарантией меньшего процента ошибок в словосочетаниях и к тому же словосочетания этого диапазона словаря могут одновременно встречаться в нескольких текстах. Аргументом в пользу включения этой частоты может также служить ее значительно меньший объем (336 423 словосочетаний (с частотой три и более) против 1 082 444 числа словосочетаний (с частотой два и более), и при этом наблюдается только незначительное снижение покрытия корпуса текстов –

с 42,5% (у словосочетаний с частотой два и более) до 34,5% (у словосочетаний с частотой три и более)).

Здесь еще нужно отметить, что исходя из нашего опыта, покрытие текстов для решения предлагаемой задачи должны быть значительно выше. Для проверки этого утверждения из части частотного словаря (с частотой три и более) был сформирован машинный концептуальный словарь и проведен эксперимент по формированию концептуальных образов документов [12] (КОДов). КОД – это совокупность выявленных в тексте значимых наименований понятий. При этом каждый элемент КОДа должен сопровождаться весовым коэффициентом, устанавливающим степень его значимости в тексте.

КОД имеет следующий вид:

$$КОД = \{Su_i \mid i \in [1, n_F]\},$$

где n_F – количество элементов в КОДе;

$Su_i = (Nc_i, w_i)$ – i-й элемент КОДа;

Nc_i – наименование понятия;

w_i – вес наименования понятия и

$$w_i = \begin{cases} (p_i + fg_i) f_i l_i, & l_i \le k_{\max}, \\ (p_i + fg_i) f_i k_{\max}, & l_i > k_{\max}, \end{cases}$$

где p_i – коэффициент, увеличивающий степень значимости наименования понятия в зависимости от его принадлежности к фамильно-именной группе, географическим названиям и т.д.;

l_i – количество слов в словосочетании, которым выражается понятие;

f_i – частота появления наименования понятия в тексте;

fg_i – коэффициент значимости понятия в предметной области (указан в словаре предметной области, иначе равен 0);

$k_{\max}$ – коэффициент, установленный опытным путем, соответствующий максимальной длине словосочетания, после которой она не должна влиять на итоговый вес наименования понятия.

Предварительный анализ сформированных КОДов показал, что в них отсутствовали некоторые значимые для анализируемого текста словосочетания. В связи с этим было решено расширить формируемый концептуальный словарь имеющимся у авторов настоящего исследования массивом англоязычных словосочетаний, общим объемом 760 тыс. Это обусловлено тем, что при добавлении к полученному частотному словарю имеющегося – точного, который состоит из проверенных специалистами значимых наименований понятий, будет получена возможность распознавать словосочетания с частотой 1 и 2, которые не попали в частотный словарь. При объединении части частотного словаря (336 тыс.) и словаря словосочетаний (760 тыс.) и исключении дублирующих словосочетаний был сформирован словарь общим объемом 920 тыс. Полученный словарь был положен в основу формируемого англоязычного концептуального словаря.

4.2 Исследование качества концептуального словаря

Эффективность функционирования процедуры концептуального анализа [6–7] зависит от качества концептуального словаря и его способности выявлять в текстах наименования понятий, выражающих их смысловое содержание. Качество концептуального словаря определяется наличием в его составе основного понятийного состава по анализируемому спектру тематических областей. Это было обеспечено разработанной технологией автоматизированного составления концептуального словаря и дополнительными мерами по расширению его состава.

Что касается исследования способности процедуры концептуального анализа выявлять в текстах наименования понятий, то для этого необходимо с помощью полученного словаря обработать несколько тестовых тематически близких документов, с приемлемым уровнем орфографических и синтаксических искажений в документах.

В качестве таких документов были выбраны пять документов, относящихся к тематике «Компьютерная лингвистика». В таблице 3 приведены статистические данные о тестовых анализируемых текстах.

Таблица 3

Статистические данные об анализируемых тестовых документах

Номер документа	Объем (в байтах)	Количество предложений	Размер КОДа (количество словосочетаний)	Количество разных слов в текстах
_0001	9972	53	44	342
_0002	21737	214	83	508
_0003	32287	287	187	1084
_0004	11458	83	52	294
_0005	20325	140	79	472

В таблице 4 приведен фрагмент результатов попарного сопоставления элементов КОДов тестовых документов. В первой колонке приведено количество совпавших в документах словосочетаний, во второй – вес совпавших словосочетаний, в третьей – % совпавших словосочетаний от общего объема словосочетаний, содержащихся в тексте.

Анализ проведенного эксперимента по оценке качества концептуального словаря и его способности выявлять в текстах наименования понятий на тестовых документах показал удовлетворительные результаты. Состав элементов КОДов соответствует смысловому содержанию документов и обеспечивает возможность сопоставления документов по их тематической близости.

Таблица 4

Результаты сравнения тестовых документов по КОДам

ТЕКСТЫ	_0004			_0005		
	Кол-во с./с.	Вес	% совп.	Кол-во с./с.	Вес	% совп.
_0001	3	37	7	9	29	8
_0002	9	64	12	11	27	5
_0003	13	91	8	9	38	3
_0004	–	–	–	78	472	45
_0005	24	105	22	–	–	–

В таблице 5 приведены совпавшие элементы КОДа обоих документов, они даны в сопровождении номеров предложений, в состав которых они входят. Из таблицы 5 видно, что в текстах _0004 и _0005 совпало 24 элемента КОДов.

Таблица 5

Совпавшие элементы КОДа текста №_0004 и текста №_0005

important challenge problem *** 00005 / 00005
computer science university of texas at austin *** 00003 / 00001
mooney department of computer science university *** 00003 / 00001
department of computer science university of texas *** 00003 / 00001
semantic analysis *** 00016 / 00009 00025 00040 00097 00102
map natural-language sentence *** 00017 / 00127
university station c0500 austin *** 00003 / 00001
database query *** 00080 / 00011 00025 00054
artificial intelligence *** 00064 00080 00081 / 00137
speech recognition *** 00040 / 00008 00041
semantic role *** 00015 / 00053 00099
individual word *** 00032 00053 / 00047
nlp system *** 00011 00013 / 00103
word and phrase *** 00030 / 00076
word sense *** 00015 / 00009
specific application *** 00020 / 00105
semantic-parser acquisition *** 00039 / 00093
past ten to fifteen year *** 00010 / 00007
parser that map *** 00017 / 00087
knowledge representation *** 00026 / 00033
knowledge engineering *** 00014 / 00007
important issue *** 00038 / 00003
robocup simulator *** 00044 / 00086
inductivelogic *** 00080 / 00058

4.3 Эксперимент по установлению тематической близости массива документов журнала «Science»

В качестве исходных данных для проведения эксперимента был использован массив англоязычных текстов журнала «Science», включающий 1835 документов по широкому спектру тематических областей. В качестве инструмента для проведения исследований был использованы разработанные в рамках настоящих исследований следующие процедуры:

1. Процедура, реализующая построение КОДа по тексту англоязычного документа. Выделение наименований понятий производилось по эталонному словарю, также ранее созданному в рамках данного исследования.

2. Процедура, реализующая сопоставление КОДа двух документов (второго по отношению к первому), визуализацию совпавших элементов КОДов, подсчет числа совпавших элементов КОДа, суммирование коэффициентов их значимости и процент совпадения КОДов второго документа по отношению к первому. Мера близости p-го и q-го документов вычислялась по следующей формуле:

$$K_{sim} = \frac{\sum_{j=1}^{n_\cap} w_{\cap j} \cdot \sum_{j=1}^{n_p} f_{pj}}{\sum_{j=1}^{n_p} w_{pj} \cdot \sum_{j=1}^{n_q} f_{qj}},$$

где $w_{\cap j}$ – j-я компонента вектора весовых коэффициентов наименований понятий, содержащихся в обоих текстах, причем веса берутся из КОД q-го текста.

w_{pj} – j-я компонента вектора весовых коэффициентов наименований понятий, содержащихся в p-м тексте.

f_{pj} – j-я компонента вектора частот наименований понятий, содержащихся в p-м тексте.

f_{qj} – j-я компонента вектора частот наименований понятий, содержащихся в q-м тексте.

$n_\cap$ – размерность вектора наименований понятий, содержащихся в обоих текстах.

n_p – размерность вектора наименований понятий, содержащихся в p-м тексте.

n_q – размерность вектора наименований понятий, содержащихся в q-м тексте.

Таблица 6

Статистические данные о выборке документов массива

	Номер документа	Объем текста (в символах)	КОД	
			Количество словосочетаний	Сумма коэффициентов значимости
1	00031	68298	569	1648
2	00049	46459	410	2100
3	00132	80932	492	1408
4	00168	84694	476	1604
5	00199	60942	520	2440
6	00212	100939	609	2988
7	00271	63027	462	2022
8	00294	35831	544	2102
9	00300	89613	500	1342
10	00334	60771	526	1836
11	00351	92928	745	3282
12	00386	91329	549	1698
13	00404	93889	679	1962
14	00419	82082	506	1778
15	00420	116712	689	3108
16	00423	74775	510	1900
17	00436	70421	506	1750
18	00492	76998	559	1962
19	00504	61636	485	1458
20	00560	48153	414	1796

Для выполнения основной задачи данного исследования – автоматического установления степени смысловой близости англоязычных научно-технических документов на основе анализа их смыслового содержания были построены КОДы для всех документов массива. В таблице 6 приведены статистические данные о выборке документов, принимающих участие в эксперименте по установлению тематической близости документов. В этой таблице указывается номер документа, его объем (в символах), количество выявленных словосочетаний и сумма коэффициентов их значимости.

На следующем этапе эксперимента было выполнено попарное сравнение КОДов каждого документа с КОДами всех документов массива. Результаты сравнения каждого документа были проанализированы и для каждого документа был выбран только один документ, с которым были наибольшие совпадения по сумме коэффициентов значимости. В таблице 7 приведены данные по результатам наилучших совпадений документов. В этой таблице указываются номер исходного документа и номер документа с наилучшим совпадением, для каждого совпадения указывается количество совпавших словосочетаний, сумма их коэффициентов значимости и процент совпадений.

Таблица 7

Результаты попарного сравнение КОД выборки с документами массива

	Исходный документ	Совпавший документ	КОД		
			Количество совпавших слово-сочетаний	Сумма коэф-фициентов значимости	Процент совпа-дений
1	00031	01628	32	129	7,8%
2	00049	01354	35	123	5,8%
3	00132	00404	36	138	9,8%
4	00168	00835	45	183	11%
5	00199	00271	27	150	6,1%
6	00212	01354	26	128	4,2%
7	00271	00199	27	150	7,4%
8	00294	00942	12	39	1,8%
9	00300	01354	64	189	14%
10	00334	01354	36	122	6,6%
11	00351	01347	24	24	6,2%
12	00353	00835	25	137	4,5%
13	00386	00942	39	114	6,7%
14	00404	00835	43	182	9,2%
15	00419	00942	39	337	18%
16	00420	01354	35	122	3,9%
17	00423	00695	30	180	9,4%
18	00436	01354	29	99	5,6%
19	00492	01543	35	152	7,7%
20	00504	01354	38	118	8,1%

5 Заключение

В проведенном исследовании показана принципиальная возможность на основе предлагаемых технологий автоматически создать декларативные средства для процедур автоматической обработки англоязычных текстов, в частности, для процедур автоматического установления степени смысловой близости англоязычных документов.

С помощью созданных в процессе проведения настоящего исследования программных и декларативных средств был проведен эксперимент по обработке массива текстов большого объема (1835 документов общим объемом 33 Мб текстовой информации).

В ходе эксперимента получены удовлетворительные результаты. Их анализ показал, что на количественные характеристики результатов обработки текстов повлияли ресурсные ограничения и принятые в связи с этим следующие допущения:

Как показал анализ частотного словаря, даже в его частотной части содержится около 30–40% ошибочных и неинформативных словосочетаний, в связи, с чем положенная в основу концептуального словаря частотная часть с частотой «три и более» (объемом 336 тыс. словосочетаний) фактически составила только 200 тыс. наименований понятий. Между тем по нашему опыту для решения аналогичных задач такой словарь должен быть на порядок больше. Такой словарь обеспечил бы более высокие параметры совпадения близких по смыслу документов.

Проведение автоматического анализа по текстам с большим количеством орфографических ошибок и синтаксических искажений также исказили параметры совпадения текстов (пример тому – более высокие результаты по совпадению документов на тестовых документах).

Для улучшения работы используемых процедур необходимо предпринять следующие шаги:

1. Необходимо выполнить в полном объеме комплекс технологических операций по созданию декларативных средств. Исходные тексты предварительно должны быть подвергнуты обработке процедурами формально-логического контроля и исправления орфографических и синтаксических ошибок. Большая часть процедур анализа и исправления текстовых искажений должны быть автоматизированы.

2. Необходимо обеспечить требуемые параметры концептуальных словарей (по объему и составу). При этом необходимо исходить из следующих рекомендаций: политематические или широкотематические словари должны быть объемом 2–3 млн наименований понятий. Тематические или узкотематические – около 1 млн Покрытие анализируемых текстов наименованиями понятий должно быть не менее 40–50%.

3. Необходимо, чтобы в состав декларативных средств были включены словари синонимов, гипонимов и гиперонимов как отдельных слов, так и словосочетаний. Объем первых должен быть не менее 60–80 тыс. слов, вторых – 300–500 тыс. словосочетаний. Это позволит с большей точностью отождествлять смысловые инварианты и повысит

значимость наименования понятия в канонической форме при построении КОДа.

4. Для создания промышленной высокопроизводительной системы обработки разноязычных документов в качестве хранилища исходных текстов, промежуточных результатов и конечных результатов необходимо использовать промышленные системы СУБД.

Литература

[1] Michael W. Berry, Malu Castellanos. Survey of Text Mining II: Clustering, Classification, and Retrieval // Springer, 2007. – 256 p.

[2] Кириченко К.М, Герасимов М.Б. Обзор методов кластеризации текстовой информации. // Труды международной конференции по компьютерной лингвистике и интеллектуальным технологиям Диалог'2001. М.: Наука, 2001.

[3] Чанышев О.Г. Метод кластеризации-классификации на основе бинарных классифицирующих таксонов // Труды II Всероссийской конференции «ЗНАНИЯ – ОНТОЛОГИИ – ТЕОРИИ» с международным участием, г. Новосибирск, 20–22 окт. 2009 г.

[4] Charu C. Aggarwal, ChengXiang Zhai. Chapter 4. A survey of text clustering algorithms // Springer, 2007. – 533 p.

[5] Киселев, М. Метод кластеризации текстов, основанный на попарной близости термов, характеризующих тексты, и его сравнение с метрическими методами кластеризации / М. Киселев // Интернет-математика 2007 : сб. работ участников конкурса науч. проектов по информ. поиску / [отв. ред. П. И. Браславский]. – Екатеринбург : Изд-во Урал. ун-та, 2007. – С. 74–83.

[6] Белоногов Г.Г. Теоретические проблемы информатики. Т. 2. Семантические проблемы информатики / под общ. ред. К.И. Курбакова. – М. : РЭА им. Г.В. Плеханова, 2008. – 342 с.

[7] Васильев В.Г., Кривенко М.П. Методы автоматизированной обработки текстов. – М. : ИПИ РАН, 2008. – 301 с.

[8] Борзых А.И., Брагина Г.А., Хорошилов А.А. Методы автоматической кластеризации документов в хранилищах научно-технической информации для решения задачи поиска плагиата в текстах документов // Информатизация и связь. – 2012. – Вып. 8.

[9] Захаров В. Н., Хорошилов А.А. Автоматическая оценка подобия тематического содержания текстов на основе сравнения их формализованных смысловых описаний // Труды XIV Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2012, г. Переславль-Залесский, Россия, 15–18 окт. 2012 г. – С. 189–195.

[10] Захаров В. Н., Хорошилов А.А. Автоматическое формирование визуального представления смыслового содержания документа // Системы и средства информатики. 2013. – Т. 23, № 1. – С. 143–158.

[11] Захаров В.Н., Хорошилов А.А. Методы решения задачи автоматического выявления заимствований в структурированных научно-технических документах на основе их семантического анализа // Труды XV Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2013, г. Ярославль, 14–17 окт. 2013 г. – С. 322–329.

[12] Хорошилов А.А. Методы автоматического установления смысловой близости документов на основе их концептуального анализа // Труды XV Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2013, г. Ярославль, 14–17 окт. 2013 г. – С. 369–376.

[13] Хорошилов А.А. / Системы обнаружения плагиата нового поколения, базирующиеся на методах концептуального анализа текстов и использовании предметно-ориентированных концептуальных словарей // Информатизация и связь. – 2013. – № 3. – С. 112–118.

[14] Viktor Zakharov, Alexey Khoroshilov / Методы решения задачи автоматического выявления заимствований в структурированных научно-технических документах на основе их семантического анализа (Semantic Methods for Solving a Problem of Automatic Detection of Plagiarism in Structured Scientific and Technical Documents) // Selected Papers of the 15th All-Russian Scientific Conference "Digital Libraries: Advanced Methods and Technologies, Digital Collections" (RCDL 2013), Yaroslavl, Russia, October 14–17, 2013. – P. 165–172.

Experience in Document Clustering on the Basis of the Method for Determining Their Thematic Similarity

Victor N. Zakharov, Alexandr A. Khoroshilov, Alexey A. Khoroshilov

The paper describes the experience in using the methods for formalized semantic description generation and thematic text content similarity estimation. The methods used in the study are based on the use of the procedures for the semantic-syntactic and conceptual analysis. These procedures provide the definition of the conceptual text content and assignment of weight characteristics to the concept names. These characteristics correspond to their semantic role and significance in the text. The software complex was created for carrying out this work, and it was tested on English-language texts.