Методы автоматического установления смысловой близости документов на основе их концептуального анализа

© А. А. Хорошилов
ЦИТиС, г. Москва
a.a.horoshilov@mail.ru

Аннотация

Приведено современное представление о смысловой структуре текстов и представлены результаты исследований проблемы автоматического установления смысловой близости документов. В процессе исследований были разработаны три модели представления смысловой структуры текста, базирующиеся на различных уровнях единиц смысла – понятиях (модель: концептуальный образ документа (КОД)), предложениях (модель: формализованная предикатно-актантная структура документа (ФПАС)) и сверхфразовых единствах (модель: текстовый фрагмент). Описаны также методы установления смысловой близости документов, в которых используются предлагаемые модели, и рассмотрены их достоинства и недостатки, а также области применения.

1 Введение

1.1 Проблемы обработки текстовой информации

Лавинообразный рост объемов текстовой информации в интернете и потребность в ее быстрой и качественной обработке привели к необходимости создания новых технологий автоматического анализа текстов. Успехи в решении этой задачи зависят, прежде всего, от достижений в изучении процессов человеческого мышления, речевого общения между людьми и умения моделировать эти процессы на компьютере. Основной проблемой, возникающей при работе с текстами документов, является трудность формализации смыслового содержания документов и, как следствие этого, трудность установления смысловой связи между различными документами. Такая проблема возникает в процессе функционирования различных систем обработки текстовой информации: в поисковых системах – при установлении смысловой близости содержания запросов и документов; в системах классификации текстов – при их распределении по классам на основе признаков сходства и различия, отражающих наиболее существенные черты смыслового содержания этих текстов; в аналитических системах – при установлении смыслового тождества или смысловой близости анализируемых документов.

Сложность этой проблемы обусловлена еще и тем, что в разных текстах одни и те же ситуации могут описываться в терминах различной степени общности и с помощью различных языковых средств. И только человек, анализирующий документы, руководствуясь своими представлениями о содержании документов и средствах выражения этого содержания и опираясь на свои профессиональные знания и опыт, в состоянии установить степень смысловой близости анализируемых документов.

Большинство систем автоматической обработки текстовой информации, функционирующих в настоящее время, не могут в полной степени решать эти проблемы. В связи с этим возникает необходимость в разработке эффективных методов автоматического анализа содержания документов и установления их смысловой близости.

1.2 Цели исследования

Целью исследований является разработка методов автоматического установления смысловой близости документов в системах автоматической обработки текстовой информации. В соответствии с указанной целью в работе поставлены следующие задачи:

- провести анализ основных подходов и методов, применяющихся при решении задачи автоматического установления смысловой близости документов;
- разработать инструментарий для решения задачи автоматического установления смысловой близости документов;
- разработать методы и алгоритмы автоматического установления смысловой близости документов на основе современных теоретических представлений о смысловой структуре текстов.
- провести экспериментальные исследования, показывающие эффективность предложенных методов.

Труды 15-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2013, Ярославль, Россия, 14-17 октября 2013 г.

1.3 Обзор методов сравнения текстов

В настоящее время для решения различных задач анализа текстовой информации используется большое число методов сравнения текстов [9 – 12], но наибольшее распространение получили такие методы, как TF, Opt Freq, Lex Rand, Log_Shingle, Megashingles, Long Sent, Descr Words. Исследованию возможностей этих методов посвящена работа [9], в которой описан широкомасштабный эксперимент по сравнительному анализу указанных методов. В этой работе ставилась задача оценить качество наиболее известных, разнообразных и эффективных с вычислительной точки зрения алгоритмов определения нечетких дубликатов. При этом предполагалось сравнивать алгоритмы по параметрам полноты и точности, а также определить их взаимную корреляцию и совместное покрытие разными сочетаниями алгоритмов исходного множества пар нечетких дубликатов. В качестве тестового массива использовалась веб-коллекция документов РОМИП (около 500 тыс. документов).

В исследуемых алгоритмах в качестве одного из параметров меры сходства документов были использованы различные текстовые фрагменты (буквенные подстроки, фиксированные последовательности значимых слов («шинглы»), частотные словари слов и т. д.), подвергнутые статистической обработке. При этом лучшие результаты по точности были у алгоритмов, базирующихся на использовании более длинных фрагментов текста. Алгоритмы, базирующиеся на более коротких фрагментах текста, обеспечивали лучшую полноту, но проигрывали в точности сравнения.

Необходимо отметить, что во всех рассмотренных алгоритмах текст рассматривается как некоторое множество, состоящее из отдельных слов. Различные операции, выполняемые в процессе поиска текстов-дубликатов, производились над словами и их цепочкам. Между тем, текст – это не множество слов и их последовательностей, и при установлении смысловой близости документов нужно сопоставлять, прежде всего, смысловые единицы текста. При этом необходимо учитывать такое явление, как вариативность форм представления в тексте одного и того же смысла.

В последнее время в работах зарубежных авторов получили широкое распространение семантические методы сравнения текстов. В [25 – 27] для выявления близких по смыслу документов (дубликатов) используется так называемый глубокий семантически ориентированный подход. В его основе лежит использование семантических сетей, которые получаются при помощи семантико-синтаксического анализатора. При этом учитываются как лексические, так и семантические отношения в тексте. При использовании этого метода были выявлены сложности при обработке неправильных и омонимичных фраз, а также отрицательных фраз.

Похожий подход используется и в работах [28 – 30]. В качестве инструмента для установления семантических отношений авторы используют электронный тезаурус WordNet. Одной из оригинальных идей, изложенных в данной работе, является то, что семантические профили слов выражаются в терминах явных (LSA), неявных (ESA) и характерных (SSA) понятий. Это решение позволяет перейти от разряженного пространства слов к более богатому и понятному пространству понятий и устанавливать отношения смысловой близости понятий. Для определения меры сходства текстов используется стандартный метод косинусов.

2 Единицы смысла языка и речи

Исследованиями смысловой структуры текстов занималось значительное число исследователей, но наибольший вклад внесли такие выдающиеся ученые, как И.А. Мельчук [20 – 21], создавший общеизвестную теорию «Смысл ⇔ Текст», и его последователи (Ю.Д. Апресян, И.М. Богуславский, Л.Л. Иомдин [22]). Согласно этой теории описание естественного языка понимается как «система правил», обеспечивающая переход от смысла к тексту («говорение», или построение текста) и от текста к смыслу («понимание», или интерпретация текста).

Несколько иной точки зрения придерживаются известные специалисты в области автоматической обработки текстовой информации – Г.Г. Белоногов и Р.С. Гиляревский [3, 18 – 19], считающие, что смысловое содержание текстов выражается с помощью единиц смысла, входящих в их состав. По их мнению, наиболее устойчивыми единицами смысла являются понятия. Они занимают центральное место в языке и речи и являются теми базовыми строительными блоками, на основе которых формируются смысловые единицы более высоких уровней. Второй по значимости единицей смысла является предложение. Из предложений формируются различного рода сверхфразовые единства, которые представляются в виде последовательностей связного текста. В связном тексте предложения выступают не изолированно друг от друга, а в тесной смысловой связи. В основе этой связи лежат мыслительные образы тех конкретных или абстрактных объектов (ситуаций, явлений), которые человек имеет в виду, когда порождает текст. Образы этих объектов имеют определенную структуру. Кроме того, они дополнительно структурируются человеком при их описании на естественном языке. Соответственно этому структурируется и текст.

Важно отметить, что в текстах одни и те же объекты и процессы могут описываться с различной степенью общности и с помощью различных языковых средств. Поэтому при решении задач автоматической смысловой обработки текстовой информации необходимо в той или иной мере учитывать такие явления, как синонимия, гипонимия (родовидовые отношения), разнообразие средств выражения межфразовых связей.

Исходя из вышесказанного, для решения поставленной задачи необходимо создать средства, позволяющие автоматически строить формализованное представление смысловой структуры текста. При

построении такого представления необходимо проанализировать текст, разделить его на предложения, выделить из него единицы смысла (наименования понятий) – слова и словосочетания, выражающие понятия. Задачу автоматического установления степени смысловой близости документов возможно выполнить путем сопоставления формализованного смыслового представления анализируемых текстов. Автоматическая формализация смыслового содержания текстов базируется на описанных ниже технологиях и процедурах автоматической обработки текстов.

3 Технологии и процедуры автоматической обработки текстов

Основным назначением технологий и процедур автоматической обработки текстовой информации является решение таких задач, как 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. Центральной процедурой систем автоматической смысловой обработки текстов является процедура их семантико-синтаксического концептуального анализа. Эта процедура реализована как процедура фразеологического концептуального анализа текстов, базирующаяся на мощном словаре наименований понятий. Базовой процедурой смысловой обработки текстов является процедура морфологического анализа.

3.1 Морфологический анализ слов

Морфологический анализ слов естественных языков предназначен для определения структуры слов и назначения им грамматических признаков, необходимых для выполнения различных процедур автоматической обработки текстовой информации, таких, например, как морфологического синтеза слов, синтаксического анализа текстов, синтаксического синтеза текстов и концептуального анализа.

Используемая нами процедура морфологического анализа базируется на оригинальных алгоритмах и методах создания машинных словарей и уникальных быстродействующих методах поиска в них.

На основе морфологического анализа строится процедура нормализации (лемматизации) слов – процедура приведения текстовых форм слов к их нормализованной форме. Обычно под нормализованной (канонической) формой слова понимается та его форма, которая традиционно указывается в словарях.

3.2 Семантико-синтаксический анализ

Семантико-синтаксический анализ проводится с целью получения формализованного представления структуры текстов – выделения в них смысловых единиц и установления связей между ними [1, 3–5, 7, 24]. В результате анализа в тексте должны быть выделены составные части, которыми являются речевые отрезки, обозначающие понятия: слова, словосочетания, фразы, сверхфразовые единства. При описании синтаксической структуры текстов в качестве одной из формализованных моделей была использована модель дерева зависимостей. Согласно этой модели каждое предложение представляется в виде дерева, в узлах которого находятся слова. Отношения непосредственной доминации визуализируются путем указания для каждого подчиненного слова (“слуги”) его подчиняющего слова (“хозяина”). При этом степень дифференциации этих отношений может быть различной, в частности, иногда достаточно установления только факта наличия смысловой связи.

Алгоритм синтаксического анализа текстов, как и множество подобных ему алгоритмов, имеет тот недостаток, что в нем в явном виде не выделяются смысловые единицы, выраженные словосочетаниями. В свою очередь смысловое содержание текстов документов выражается с помощью единиц смысла – понятий и связей между ними. Г.Г. Белоногов [3, 4] определяет понятие как социально значимый мыслительный образ, за которым в языке закреплено его наименование в виде отдельного слова или, значительно чаще, устойчивого фразеологического словосочетания. Под устойчивыми фразеологическими словосочетаниями мы будем понимать не только идиоматические выражения и терминологические словосочетания, но и любые повторяющиеся отрезки связных текстов, для их выделения применяется процедура концептуального анализа.

3.3 Концептуальный анализ текстов

Концептуальный анализ текстов предназначен для определения смысловой структуры текстов, выявления их понятийного (концептуального) состава и установления смысловых связей между наименованиями понятий [23]. В более узком понимании концептуальный анализ можно рассматривать как процедуру выявления наименований понятий в текстах. Это сложная задача, и ее невозможно решить только путем анализа синтаксической структуры текстов; необходимо также привлекать семантические признаки.

3.4 Технология составления частотных словарей по корпусу текстов

Как было указано выше, 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 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 него исключаются некорректные слова и словосочетания. И, наконец, частотная часть словаря подвергается лингвистической обработке, в результате которой исключается малоинформативная и некорректная лексика.

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

- формально-логический контроль исходных текстов с целью обнаружения и исправления орфографических и синтаксических ошибок в исходных текстах;
- членение исходного текста на отдельные слова (по пробелам и разделительным знакам между ними);
- морфологический анализ слов корпуса текстов;
- членение корпуса текстов на предложения;
- семантико-синтаксического анализ текстов;
- приближенный концептуальный анализ текстов;
- выделение наименований понятий;
- автоматическое приведение наименований понятий к их канонической форме;
- формирование частотного словаря наименований понятий;
- лингвистический анализ частотного словаря наименований понятий (исключение ошибочной и малоинформативной лексики);
- формирование машинного представления концептуального словаря.

4 Модели представления смыслового содержания документов

При построении понятийной модели текста необходимо опираться на современные теоретические представления о его смысловой структуре. В соответствии с теоретическими представлениями, положенными в основу наших исследований, смысловое содержание текстов документов выражается с помощью единиц смысла, входящих в их состав: понятий, предложений и различного рода сверхфразовых единств. В соответствии с этими представлениями было разработано несколько моделей представления смысловой структуры текстов документов.

4.1 Модель 1: концептуальный образ документа (КОД)

Первая из предлагаемых моделей базируется на наиболее устойчивых единицах смысла – понятиях. Именно понятия являются теми элементарными мыслительными образами, на основе которых строится более сложный мыслительный образ, соответствующий анализируемому тексту. Совокупность выявленных в тексте наименований понятий будем называть концептуальным образом документа (КОДом). При этом каждый элемент КОДа должен сопровождаться весовым коэффициентом, устанавливающим степень его значимости в тексте.

Концептуальные образы документов создаются на основе их автоматического концептуального анализа – выделения в них наиболее значимых наименований понятий. Для этого необходимо располагать достаточно мощным словарем наименований понятий, включающим в свой состав наиболее информативные понятия поискового массива. Это достигается путем составления по текстам поискового массива частотного словаря наименований понятий и его последующего редактирования.

4.2 Модель 2: формализованная предикатно-актантная структура документа (ФПАС)

Следующая модель опирается на вторую по значимости единицу смысла – предложение. Из предложений формируются различного рода сверхфразовые единства, которые представляются в виде последовательностей предложений (связного текста). Основной чертой предложений является их предикативность – то их свойство, что в них утверждается наличие у объектов определенных признаков и их отношений [2, 6, 8]. Свойством предикативности обладают и высказывания, формулируемые на формализованных языках. Это позволяет сделать вывод, что в основе и предложений на естественном языке, и формализованных логических высказываний лежит предикатно-актантная структура, компонентами которой являются понятия-предикаты (признаки, отношения) и понятия-актанты, выступающие в роли описываемых объектов. В естественных и формализованных языках предикатно-актантные структуры являются теми смысловыми инвариантами, которые позволяют осуществлять автоматический перевод текстов с естественных языков на формализованные и с формализованных на естественные.

Формализованные предикатно-актантные структуры (ФПАС) создаются на основе автоматического концептуального анализа текстов – выделения в них наиболее значимых наименований понятий и установления связей между ними. Для этого также необходимо располагать словарем наименований понятий-актантов и словарем наименований понятий-предикатов. ФПАС состоит из совокупности наименований понятий-актантов, связанных между собой смысловыми отношениями-предикатами. Эта модель может быть представлена в виде графа, вершинами которого являются актанты, а дугами – отношения-предикаты.

4.3 Модель 3: сверхфразовое единство

Третья модель основывается на формализации содержания следующего уровня единиц смысла – сверхфразового единства, в роли которого выступают текстовые фрагменты, состоящие из двух и более предложений и связанные общей темой. В качестве такой единицы смысла часто выступает текст документа.

Общеизвестно, что используемые в текстах слова и словосочетания проявляют свои основные свойства, только будучи «текстово-связанными», т. е. тогда, когда они образуют тексты и передают их содержание. Значимые объекты, отображаемые в речи последовательностью слов, включены в предложения (а через них – в тексты) и, кроме того, в «контекст» отображаемой ситуации. При этом семантика слов в тексте (их значение и смысл) может значительно отличаться от семантики изолированных слов, поскольку только в развернутом высказывании слово получает свои «реальные» значение и осмысление. Поэтому при определении смысла-содержания языковых единиц одного уровня требуется обращение к единицам более высокого уровня. Текст выступает в данном случае предельной (высшей) единицей общения на знаковом уровне. Все это делает необходимым при определении семантики (смысловой, содержательной стороны) речи всегда анализировать ее «текстовой континуум» [17].

Формализованное представление текстовых фрагментов можно обеспечить путем их пословной нормализации и представления их в виде упорядоченной совокупности значимых слов текстового фрагмента (предложения или различного рода сверхфразовых единств). При этом для последующего анализа должна сохраняться информация о порядке следования слов в текстовом фрагменте.

5 Методы установления смысловой близости документов

При разработке методов автоматического установления смысловой близости документов необходимо опираться на формализованные представления их смысловой структуры, обеспечивающие возможность сравнения смыслового содержания документов. При этом такие представления могут описывать смысловое содержание текстов с различной степенью подробности и общности, в зависимости от того, какая модель положена в основу конкретного формализованного представления текстов.

5.1 Метод установления смысловой близости документов по КОДам

Идея автоматического распознавания смысловой близости документов по их КОДам заключается в том, что концептуальный образ анализируемого документа (его КОД) попарно сравнивается с концептуальными образами (КОДами) документов массива [13, 14]. При этом для каждой пары сравниваемых КОДов определяется коэффициент их смысловой близости. В упрощенной постановке задачи этот коэффициент можно было бы положить равным отношению суммы весов наименований понятий КОДа анализируемого документа, совпавших с наименованиями понятий КОДа из массива, к сумме весов всех наименований понятий КОДа документа.

Однако такой подход является упрощенным, следовательно, и некорректным потому, что обычно в разных текстах одни и те же объекты и процессы могут описываться с различной степенью обобщения и с помощью различных языковых средств. Поэтому при решении задач автоматического распознавания смысловой близости документов необходимо в той или иной мере учитывать такие явления, как словоизменение, словообразование, синонимию, гипонимию (родо-видовые отношения), разнообразие средств выражения межфразовых связей.

Явление словоизменения может быть учтено путем применения процедуры автоматического морфологического анализа слов и отождествления различных форм слов по их основам. Для учета явления словообразования можно использовать словарь словообразовательных вариантов слов. Для учета явлений синонимии и гипонимии необходимо использовать словарь синонимов, гипонимов (более узких по объему понятий) и гиперонимов (более широких по объему понятий) как на уровне отдельных слов, так и на уровне фразеологических словосочетаний.

Сложнее дело обстоит с учетом таких явлений, как вариации обозначений одних и тех же понятий в связных текстах. Дело в том, что в связном тексте наименование понятия, выраженное некоторым фразеологическим словосочетанием, может быть сначала представлено в своей исходной форме, но затем, в последующих предложениях, – в сокращенных вариантах. Оно может быть также заменено на наименование родового понятия или на местоимение.

Отрицательное влияние этих явлений на процесс распознавания смысловой близости документов частично сглаживается тем, что различные варианты наименований одних и тех же понятий будут включены в частотный словарь наименований понятий при его составлении и, как следствие, в концептуальные образы документов.

При сопоставлении наименований понятий концептуальных образов анализируемого документа и документов из массива необходимо в первую очередь учитывать явления словоизменения, синонимии и гипонимии, так как эти явления оказывают наибольшее влияние на полноту распознавания смысловой близости документов. При этом информацию о различных вариантах синонимов, гипонимов и гиперонимов слов и словосочетаний следует вносить в концептуальный образ анализируемого документа, а не в концептуальные образы документов массива, так как во втором случае это привело бы к существенному увеличению поискового массива.

При сопоставлении однословных наименований понятий они считаются связанными по смыслу, если выполняется следующее условие: либо они полностью совпадают, либо являются синонимами, либо

находятся в родо-видовых отношениях. При сопоставлении фразеологических словосочетаний к ним предъявляются те же требования, что и к однословным наименованиям понятий, но если указанное выше условие не выполняется, то предпринимается попытка пословного сопоставления словосочетаний. В этом случае словосочетания считаются связанными по смыслу, если в обогащенном КОДе документа для каждого слова словосочетания из КОДа документа массива находится хотя бы один синоним, гипоним или гипероним.

По окончании процесса сопоставления всех наименований понятий КОДа анализируемого документа и КОДа документа из массива определяется коэффициент смысловой близости этих документов. При этом производится суммирование «весов» исходных наименований понятий КОДа документа, связанных по смыслу с наименованиями понятий КОДа документа из массива, и полученная сумма делится на сумму весов всех исходных наименований понятий КОДа документа. Документы считаются связанными по смыслу, если коэффициент их смысловой близости превышает заданный порог значимости.

5.2 Метод установления смысловой близости документов по ФПАСам

Идея автоматического распознавания смысловой близости документов по ФПАСам заключается в попарном определении меры подобия графов ФПАСа анализируемого документа и ФПАСов документов массива [15]. В данном исследовании для сопоставления графов будем пользоваться методом, описанным в работе [16]. Этот метод сравнения используется для поиска изоморфных пересечений двух графов. Как утверждают авторы, метод основан на построении графов, по своей структуре сходных с нейронными сетями и названных пирамидами. При данном подходе сначала строится пирамида на основе одного графа, затем на основе второго графа строится пирамида, сходная по структуре с первой. Каждой вершине второй пирамиды соответствует подграф второго графа, изоморфный (гомоморфный) подграфу первого графа. Построение пирамид проводится за полиномиальное время, что делает данный метод применимым в данной задаче.

При решении задачи построения ФПАСа документа необходимо также решить проблему вариативности форм представления одних и тех же наименований понятий-актантов. Поэтому различные формы представления наименований предварительно должны быть приведены к их канонической форме. Понятия-предикаты также нужно привести к их типизированным формам, то есть формам, обозначающим одинаковые действия, процессы, признаки и др.

Аналогично предыдущему методу для каждого наименования понятия-актанта и понятия-предиката должны быть назначены соответствующие весовые коэффициенты.

5.3 Метод установления смысловой близости документов по текстовым фрагментам

Метод установления смысловой близости документов по текстовым фрагментам базируется на предположении, что если в двух текстах имеются такие текстовые фрагменты, в которых содержатся одинаковые значимые слова, и их доля превышает заданный порог, то эти фрагменты находятся в смысловой связи. Степень их смысловой близости будет определяться отношением количества совпавших значимых слов к их общему числу. С целью обеспечения установления смыслового тождества значимых слов на уровне словоизменения все слова текстового фрагмента должны быть нормализованы.

Для повышения точности установления смысловой близости таких фрагментов необходим дополнительный анализ как отдельных слов, так и контактно-расположенных последовательностей значимых слов на возможность того, что они являются наименованиями понятий или включают их. Этот анализ также необходимо проводить с учетом вариативности форм представления понятий в текстах.

В связи с вышесказанным для реализации количественных оценок с учетом сопоставления обнаруженных наименований понятий в этих текстовых фрагментах необходимо также установить весовые коэффициенты смысловой значимости наименований понятий. Тогда степень смысловой близости текстовых фрагментов будет определяться не отношением количества совпавших слов к их общему числу, а отношением суммы весов совпавших наименований понятий к общему весу текстового фрагмента.

6. Эксперименты по установлению смысловой близости документов

6.1 Описание планируемых экспериментов

На данном этапе работы были проведены предварительные эксперименты, позволяющие опробовать методы, предложенные в статье. Например, в работе [15] описан эксперимент по установлению заимствований в научно-технических документах с помощью метода, приведенного в пункте 5.3. Для проведения исследований создается программный комплекс, позволяющий сравнивать документы всеми предложенными методами. В этом программном комплексе будут реализованы методы сравнения документов, которые описаны в [9]. Эксперимент будет проводиться на массиве документов общественно-политической тематики, содержащихся в системе «Мониторинг СМИ» (www.мониторингсми.рф).

6.2 Описание методики оценки эффективности предложенных методов

Для оценки эффективности предлагаемых методов предлагается провести две группы экспериментов.

Первая заключается в том, что экспертами со-

здаются группы документов, похожих на документ-образец, которые ранжируются по степени смысловой близости к данному документу. Затем в автоматическом режиме ранжирование будет произведено с помощью программного обеспечения, реализующего методы, предложенные в настоящей статье. После этого будет проведено сравнение результатов работы экспертов и системы.

Вторая группа экспериментов заключается в том, что экспертами выбирается в массиве документов несколько документов-образцов, после этого подбираются документы, похожие на них по смысловому содержанию. Затем с помощью программного обеспечения документы в автоматическом режиме будут соотноситься с документами-образцами. После этого будет проведено сравнение результатов работы экспертов и системы.

7 Заключение

Анализ эффективности предлагаемых методов показал следующее.

Метод установления смысловой близости документов по их КОДам наилучшим образом подходит для предварительной оценки содержания текстов на наличие в них схожих объектов или ситуаций. Достоинством этого метода является простота реализации, быстродействие и применимость к любой структуре текстов. Данный метод показал большую эффективность при кластеризации документов по документу-образцу. Недостатком является необходимость предварительного создания и постоянного ведения эталонных словарей наименований понятий. Данный метод показал свою эффективность при использовании его в промышленной системе «Мониторинг СМИ» (СКЦ РОСАТОМ).

Метод установления смысловой близости документов по их ФПАСам предназначен для сравнения смысловой структуры текстов с учетом взаимосвязей объектов. Достоинством этого метода является точность сравнения смысловой структуры документов, применимость к любой структуре текстов. Недостатком являются большая сложность и трудоемкость процесса. Данный метод находится в стадии разработки.

Метод установления смысловой близости документов по их нормализованным текстовым фрагментам применяется в проекте ЕСУ НИОКР (ЦИТиС) для выявления заимствований в текстах документов. Достоинством этого метода являются высокое быстродействие и точность выявления одинаковых текстовых фрагментов в текстах документов. Недостатком является относительно низкая эффективность сравнения близких по смыслу документов, имеющих различные лексический состав и синтаксическую структуру.

Следующим этапом работы будут завершение реализации всех предложенных методов и создание программного обеспечения, позволяющего производить сравнение текстов тремя различными способами. Также необходимо провести эксперименты по сравнению текстов, описанные в п. 6. После завершения данных экспериментов планируются встраивание реализованных процедур в различные поисковые системы и другие системы обработки текстовой информации, а также построение собственной системы установления заимствований в текстах.

Литература

[1] Кузнецов И.П. Механизмы обработки семантической информации. – М.: Наука, 1978. – 175 с.

[2] Осипов Г.С. Приобретение знаний интеллектуальными системами: Основы теории и технологии. – М.: Наука. Физматлит, 1997. – 112 с.

[3] Белоногов Г.Г., Гиляревский Р.С. и др. Развитие систем автоматической обработки текстовой информации// Нейрокомпьютеры: разработка, применение. – 2010, № 8.

[4] Белоногов Г.Г. Теоретические проблемы информатики, Том 2. Семантические проблемы информатики. Под общей редакцией К.И. Курбакова. – М.: РЭА им. Г.В. Плеханова, 2008. – 342 с.

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

[6] Крейнес М.Г. Обеспечение активности содержания многоязычия текстовых документов: технология КЛЮЧИ ОТ ТЕКСТА. – Информационное общество. – 2000, Вып. 2, 241 с.

[7] Соссюр Фердинанд де. Курс общей лингвистики. – М.: Прогресс, 1977. – 370 с.

[8] Чугреев В.Л. Модель структурного представления текстовой информации и метод ее тематического анализа на основе частотно-контекстной классификации// Дис. … канд. техн. наук. – Санкт-Петербург, 2003. – 185 с.

[9] Зеленков Ю.Г., Сегалович И.В. Сравнительный анализ методов определения нечетких дубликатов для Web-документов // Труды 9 ой Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL’2007, Переславль, Россия, 2007. – Том 1, С. 166-174.

[10] Manber U.. Finding Similar Files in a Large File System. Winter USENIX Technical Conference, 1994.

[11] Broder A., Glassman S., Manasse M., Zweig G. Syntactic clustering of the Web// Proc. of the 6th Int.l World Wide Web Conf., April 1997.

[12] Park S.-T., Pennock D., Lee Giles C., Krovetz R. Analysis of Lexical Signatures for Finding Lost or Related Documents, SIGIR’02, August 11 – 15, 2002, Tampere, Finland.

[13] 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.

[14] Захаров В.Н., Хорошилов А.А. Автоматическая оценка подобия тематического содержа-

ния текстов на основе сравнения их формализованных смысловых описаний // Труды XIV-ой Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2012, г. Переславль-Залесский, Россия, 15 – 18 октября 2012 г.

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

[16] Агарков А.В. Метод сравнения двух графов за полиномиальное время// Научно-теоретический журнал «Искусственный интеллект». – 2003. – №4.

[17] Ковшиков В.А., Пухов В.П. Психолингвистика. Теория речевой деятельности. – М.: Московский психологосоциальный институт, Воронеж: НПО «МОДЭК», 2001. – 432 с.

[18] Белоногов Г.Г., Гиляровский Р.С. и др. Проблемы автоматической смысловой обработки текстовой информации// Научно-техническая информация. Сер. 2. Информационные процессы и системы. – М.: ВИНИТИ РАН, 2012. – № 11.

[19] Белоногов Г.Г., Гиляревский Р.С. и др. О путях повышения качества поиска текстовой информации в системе Интернет// Научно-техническая информация. Серия 2. – М.: ВИНИТИ РАН, 2013. – № 8.

[20] Мельчук И.А. Опыт теории лингвистических моделей «Смысл ⇔ Текст». – М., 1974 (2-е изд., 1999).

[21] Мельчук И.А. Русский язык в модели «Смысл ⇔ Текст». – Москва – Вена, 1995.

[22] Апресян Ю.Д., Богуславский И.М., Иомдин Л.Л. и др. Лингвистическое обеспечение системы ЭТАП-2. – М.: Наука, 1989.

[23] Белоногов Г.Г., Быстров И.И. и др. Автоматический концептуальный анализ текстов. // Научно-техническая информация. Сер. 2. – М.: ВИНИТИ, 2002. – № 10.

[24] Звегинцев В.А. Предложение и его отношение к языку и речи. – М.: Изд-во Московского университета, 1976.

[25] Hartrumpf Sven, vor der Brück T., Eichhorn Ch. Detecting duplicates with shallow and parser-based methods// In Proc. of the 6th Int. Conf. on Natural Language Processing and Knowledge Engineering (NLPKE), Beijing, China, 2010. – P. 142-149.

[26] vor der Brück T., Hartrumpf S. A readability checker based on deep semantic indicators// In Human Language Technology. Challenges of the Information Society (edited by Vetulani, Zygmunt and Hans Uszkoreit). – 2009. – V. 5603 of Lecture Notes in Computer Science (LNCS). – P. 232-244. Berlin, Germany: Springer.

[27] Hartrumpf S., vor der Brück T., Eichhorn Ch. Semantic Duplicate Identification with Parsing and Machine Learning. – TSD 2010. – P. 84-92.

[28] Banea C., Hassan S., Mohler M., Mihalcea R. UNT: A Supervised Synergistic Approach to Semantic Text Similarity// Proc. of the Sixth Int. Workshop on Semantic Evaluation SemEval, 2012.

[29] Hassan S., Mihalcea R. Measuring semantic relatedness using salient encyclopedic concepts// Artificial Intelligence, Special Issue, 2011.

[30] Mohler M., Mihalcea R. Text-to-text semantic similarity for automatic short answer grading// In Proc. of the European Association for Computational Linguistics (EACL 2009), Athens, Greece.

Methods for automatic determination of semantic proximity of documents on the basis of their conceptual analysis

Alexey A. Khoroshilov

This paper presents the modern concept of the semantic structure of texts and the results of research on the problem of automatic determination of the semantic proximity of documents. Three models for representation of semantic structure of the text were developed in the process of research. These models are based on different levels of units of meaning - concepts (model: conceptual pattern of the document), sentence (model: formalized predicate-actantial structure of the document) and super-phrasal unities (model: text fragment). The paper also describes the methods for determination of the semantic proximity of documents that use the proposed models and discusses their advantages and disadvantages, as well as the fields of their application.