Методы решения задачи автоматического выявления заимствований в структурированных научно-технических документах на основе их семантического анализа

© В.Н. Захаров
ИПИ РАН

© А. А. Хорошилов
ЦИТиС

Москва

vzakharov@ipiran.ru

a.a.horoshilov@mail.ru

Аннотация

В работе излагаются методы выявления заимствований в структурированных научно-технических документах, базирующиеся на семантическом анализе текстов. Существующие системы позволяют устанавливать заимствования только в случаях, если эти заимствования производятся путем копирования фрагментов текста без его изменения или с незначительными изменениями его структуры или лексического состава. Использование семантических методов анализа текстов позволяет выявить смысловую структуру текста и распознавать более сложные случаи преднамеренного изменения заимствованных текстов, например, установить случаи замены слов или словосочетаний их смысловыми инвариантами, изменения разбиения текста на предложения, перестановка фрагментов текста. Также в данной работе обосновывается необходимость выявлять в текстах и исключать из рассмотрения текстовые фрагменты, относящиеся к описанию структурных элементов документов (например, стандартные для всех документов заголовки).

1 Проблема выявления заимствований в текстах документов

1.1 Введение

В соответствии с Гражданским кодексом Российской Федерации государство обязано защищать объекты авторских прав от различных нарушений в этой сфере. Одним из наиболее частых нарушений является умышленное присвоение авторства на чужие результаты интеллектуальной деятельности, представленные в виде произведений науки, литературы или искусства. Часто такое присвоение авторства осуществляется путем заимствования чужого произведения или его части. Борьба с этими нарушениями авторских прав ведется постоянно, но только сейчас предпринимаются попытки использования средств автоматизации для установления фактов таких нарушений.

1.2 Анализ существующих средств установления заимствований в документах

В качестве одного из средств поиска заимствований можно рассматривать Интернет-сервис AntiPlagiat.ru [14-16], который предлагает набор услуг, реализующих технологию проверки текстовых документов на наличие заимствований. Проверка документа выполняется путем его загрузки на сервер системы, сопоставления текста этого документа с базой данных системы и определения степени уникальности текста. При этом система выдаёт все ссылки на источники, из которых были заимствованы тексты. База данных системы «Антиплагиат» включает как открытые источники сети Интернет, так и закрытые, например полнотекстовую базу Электронной библиотеки диссертаций Российской государственной библиотеки (ЭБД РГБ).

Анализируя эту систему и ряд подобных систем [13-16] нельзя не отметить присущие им органические недостатки. Так, в частности, такие системы в значительной степени ориентированы на установление фактов прямого заимствования. При этом, системой могут выявляться заимствования, в которых была произведена незначительная замена отдельных слов на их синонимы, а также были произведены различные преднамеренные форматные искажения. Но более существенные изменения заимствованного текста, заключающиеся в расширенном использовании синонимов слов и словосочетаний, добавлении или удалении слов,

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

разбиении или объединении предложений, система не устанавливает, и такие тексты часто определяется как оригинальные.

Такая ситуация обусловлена использованием упрощенной модели представления смыслового содержания анализируемых текстов предложения или более крупного фрагмента текста. Этого вполне достаточно для выявления близких по лексическому составу и синтаксической структуре предложений или фрагментов текста.

Между тем, как утверждается в работах [3,4], связный текст это не набор отдельных предложений. Предложения выступают в тексте 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 он 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. При прочтении текста у читателя, как и у автора текста, возникнет определенный мыслительный образ. Описание этого мыслительный образа выполнено на основе применения различных языковых средств его выражения. Иными словами, одни и те же результаты интеллектуальной деятельности могут быть описаны с использованием разных форм представления смыслового содержания этого текста. При этом лексический состав и количество предложений текста может быть различными. В этом случае для выявления более сложных изменений заимствованного текста необходимо использовать более адекватную модель представления смысловой структуры текстов.

В последнее время в работах отечественных и зарубежных авторов получают широкое распространение семантические методы сравнения текстов. В работе [20] в контексте описания оригинального решения задачи кластеризации рассматривается метод определение пар текстов с максимальной тематической близостью. Данный способ занимает промежуточное положение между синтаксическими и геометрическими методами. Каждый текст представляется набором лексем, которым поставлена в соответствие числовая характеристика их тематической важности (вес) в этом тексте. В дальнейшем определяется не геометрическое расстояние между текстами, а асимметричная близость между i-ым j-ым документами, определяемая как сумма несимметричных сумм весов лексем пересечения: веса сначала суммируются отдельно по i-ому и j-ому документам.

Таким образом, в работе основной акцент делается на правильность определения тематической роли лексемы.

В работе [21] для выявления близких по смыслу документов (дубликатов) используется так называемый глубокий семантически ориентированный подход. В основе данного метода 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 работе [22]. 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. Это позволяет устанавливать отношения смысловой близости понятий. Для определения меры сходства текстов используется стандартный метод косинусов.

1.3 Теоретическое обоснование необходимости создания нового поколения систем автоматического выявления заимствований

В соответствии с современными теоретическими представлениями в языке и речи наиболее информативными и наиболее устойчивыми единицами смысла являются понятия [3-4,7]. С их помощью описывается смысловое содержание текстов, и именно они являются теми базовыми строительными блоками, на основе которых формируются смысловые единицы более высоких уровней, в частности, предложения. При установлении смысловой близости документов нужно сопоставлять, прежде всего, смысловые единицы текста – понятия, выраженные словосочетаниями. При этом, необходимо учитывать такое явление как вариативность форм представления в тексте одного и того же смысла. Именно это явление в вышеупомянутых системах полностью игнорировалось. Поэтому алгоритмы установления смысловой близости документов должны базироваться на современных процедурах семантико-синтаксического и концептуального анализа. Использование таких процедур позволит выявить не только прямое заимствование, но и установить семантическое тождество документов, отличающихся лексическим составом текста и семантико-синтаксической структурой, но имеющих тождественное смысловое содержание[1-6,8].

Такие системы, базирующиеся на предлагаемой модели представления смыслового содержания текстов, можно отнести к системам выявления заимствований следующего поколения. Эти системы смогут выполнять наряду с функцией выявления смыслового заимствования, также функцию автоматической экспертизы документов на их научную новизну.

В процессе разработки таких систем необходимо разработать концептуальную модель предметных областей, выявить их понятийный и терминологический состав, установить систему смысловых связей между наименованиями понятий и разработать комплекс программных средств автоматического анализа смысловой структуры документов.

В наших исследованиях основной упор делается на семантический анализ содержания научно-технических документов, поэтому необходимо еще учитывать то, что такие документы имеют сложную структуру, некоторые элементы которой повторяются в большинстве исследуемых текстов. Эти элементы имеют различную степень значимости с точки зрения установления заимствований, и многие из них необходимо исключить из рассмотрения или придать меньшую степень значимости.

2 Семантические методы выявления заимствований в текстах документов

2.1 Принципы построения процедур автоматического выявления заимствований

Для проведения исследований с целью разработки семантических методов выявления заимствований было создано программное обеспечение, позволяющее сравнивать тексты документов между собой. Это программное обеспечение базировалось на ранее созданных нами процедурах упрощенного семантико-синтаксического и концептуального анализа [17-18], отличающихся более высокой скоростью обработки текстов.

В общем случае процесс выявления заимствований с использованием упрощенных семантических методов делится на несколько этапов:

1. На первом этапе в заранее обусловленной выборке текстов необходимо установить близкие по их смысловому содержанию тексты. Установление смысловой близости документов выполняется путем сравнения их формализованных смысловых описаний (ФСОД).
2. На втором этапе необходимо установить какие именно фрагменты текстов в наибольшей степени совпадают по своему содержанию с фрагментами других текстов. Для этого в анализируемых текстах устанавливаются местоположения всех совпавших элементов ФСОД (слов и словосочетаний) и выбираются те текстовые фрагменты, в которых содержится максимальное число элементов ФСОД. Таких текстовых фрагментов может быть несколько.
3. На третьем этапе необходимо установить совпадения более мелких фрагментов текстов – одного или контактно расположенных предложений. Для этого полученные фрагменты с помощью процедур семантико-синтаксического и концептуального анализа расчленяются на предложения и в них выделяются наименования понятий и устанавливаются связи между ними. Результаты этого анализа можно представить в виде таблицы связей наименований понятий или в виде графа.
4. В случае необходимости выявления преднамеренной замены наименований понятий на их смысловые инварианты, необходимо по словарю синонимичных фразеологических словосочетаний произвести во всех текстовых фрагментах автоматическую замену исходных слов и словосочетаний на их заранее установленные канонические инварианты. Далее анализ выполняется в соответствие с п.3.
5. На четвертом этапе выполняется процесс выявления заимствований, который заключается в сопоставлении полученных представлений смысловой структуры предложений и принятия решения по каждому случаю в зависимости от полученных результатов совпадения этих представлений, как в пределах конкретного предложения, так и в пределах смыслового фрагмента.

2.2 Программное обеспечение системы автоматического выявления заимствований

На основе приведённых выше принципов была создана процедура сравнения документов (см. Рис. 1). Данная процедура позволяет получить следующие характеристики для проверяемых текстов:

1. Процент совпадения первого текста со вторым.
2. Количество предложений в первом и во втором текстах.
3. Количество совпавших предложений в текстах.

Также для каждого совпавшего предложения мы получаем следующие параметры (см. таблицу 1):

1. Номер каждого из сравниваемых предложений в соответствующих текстах.
2. Вектор соответствий слов в предложениях, в котором “1” показывает, что слово входит в состав совпавшего словосочетания, а “0” означает, что слово не входит в состав совпавших словосочетаний, или оно не является значимым.
3. Процент совпавших слов.
4. Текст предложения.

Таблица 1

Параметры сравнения двух предложений

Наименование текста	Номер предложения в тексте	Вектор соответствий в сравниваемых предложениях	Процент совпадения текстовых фрагментов	Текст предложения
0220xxxxx84	505	111111110110101110	77	Изготовленный корпус оклеивается несколькими слоями стеклоткани суммарной толщиной до 5 мм, зачищается, шпаклюется синтетической шпаклёвкой.
0220xxxxx04	725	11111111011010111000000 00	58	Изготовленный корпус оклеивается несколькими слоями стеклоткани суммарной толщиной до 5 мм, зачищается, шпаклюется синтетической шпаклёвкой (рисунки 4.15, 4.16).

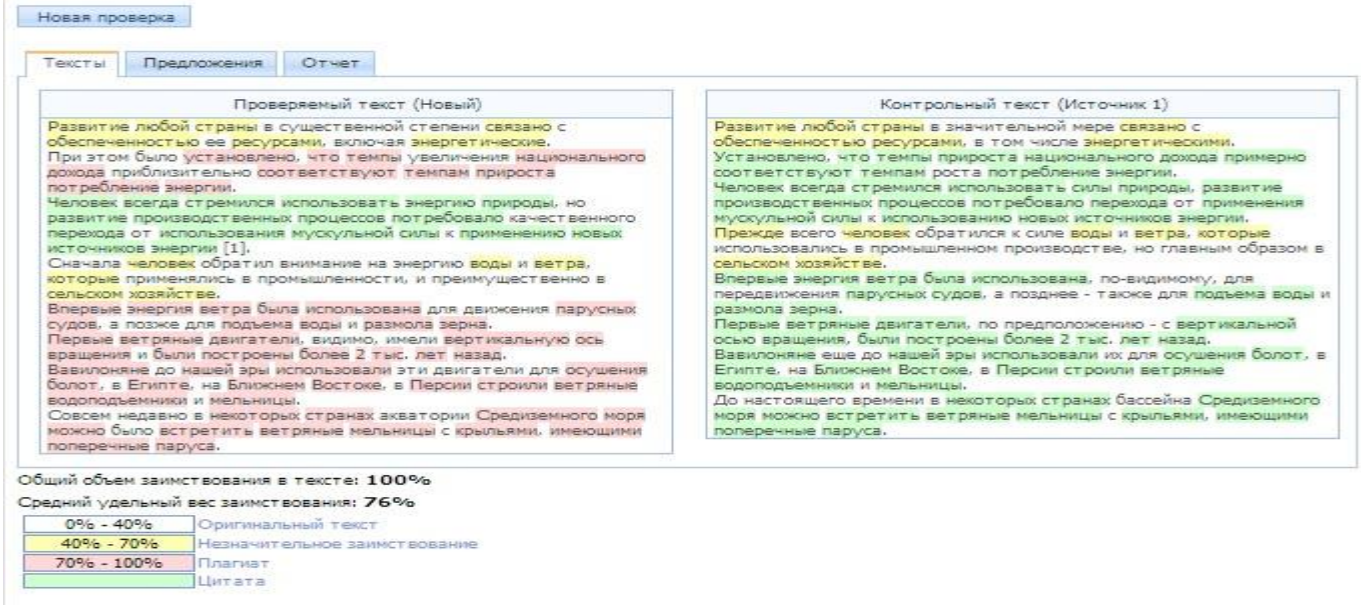

Рис. 1 Пример работы программы

3 Установление заимствований в структурированных научно-технических документах

3.1 Исходные данные

Для проведения исследования была подготовлена подборка отчетов по НИОКР, содержащихся в фонде ЕСУ НИОКР. Данные документы принадлежали к рубрике 55.45 “Судостроение” ГРНТИ за период с 2007 г. по настоящее время, всего 187 документов. Полные тексты найденных документов были обработаны программным обеспечением, описанным выше. Полученные результаты позволили выявить группы документов, в которых есть совпадающие фрагменты текста. Результаты сравнения документов, членов одной из таких групп, представлены в таблице 2. В таблицу вносились документы, в которых было больше 1.5 % совпавших предложений. Более подробные результаты сравнения данной группы документов с документом 0220xxxxx84 представлены в таблице 3. В данном случае брались только предложения, совпавшие на 90 и более процентов.

3.2 Анализ результатов эксперимента

При подробном рассмотрении результатов было установлено, что анализируемые отчеты являются различными этапами одной и той же работы и содержат общие для всей этой работы фрагменты текстов, например в текстах 0220xxxxx84 и 0220xxxxx05 есть фрагмент: «Цель работы в целом: разработка теоретических основ и экспериментальная проверка вопросов проектирования…». Естественно это не является заимствованием. Большинство таких повторов текстовых фрагментов вызвано специфической структурой документа-отчета. Такие случаи необходимо исключать из рассмотрения. В связи с этим возникает необходимость производить анализ структуры документов на основе ранее

Таблица 2

Сравнение группы из четырех документов (результатом является % совпадения)

Наименование документа	0220xxxxx84	0220xxxxx04	0220xxxxx05	0220xxxxx18
0220xxxxx84	-	8.1	4.6	10.5
0220xxxxx04	3.4	-	6.5	1.6
0220xxxxx05	2.6	8.6	-	3.3
0220xxxxx18	8.2	2.9	4.7	-

разработанного словаря структурных элементов документов.

3.3 Членение документа на смысловые фрагменты

Для обеспечения задачи установления смысловой структуры текста необходимо решить задачу разделения текста на его смысловые фрагменты, т.е. текстовые фрагменты, в которых описывается одна и та же ситуация или объект (описание образа объекта). Эта задача очень сложная, поскольку в тексте границы таких фрагментов формально не обозначены. Средствами семантико-синтаксического анализа возможно в тексте выделить только предложения. Но они, как правило, являются только частью смысловых фрагментов. Но если исходить, как было выше сказано, из того, что предложения выступают в тексте не изолированно друг от друга, а в тесной смысловой связи и в основе этой связи лежат мыслительные образы тех конкретных или абстрактных объектов (ситуаций, явлений), которые человек имеет в виду, когда он порождает текст, а также принимать во внимание, что тексты дополнительно структурируются человеком при их описании на естественном языке, то можно предположить, что описания образов этих объектов имеют определенную структуру и в тексте эта структура обозначена в виде абзаца. В тех случаях, когда текст не структурирован, такое членения необходимо выполнить автоматически, исходя из вышеприведенных рассуждений.

Для решения задачи автоматического выявления в тексте смысловых фрагментов, необходимо выполнить его семантико-синтаксический и концептуальный анализ. При этом в тексты будут установлены границы предложений, выявлены наименования понятий и установлены смысловые связи между ними. После построения таблицы связей можно по ней выявить все контактно расположенные предложения, в которых содержатся наименования понятий, связанные с ключевыми наименованиями понятий. Границами таких смысловых блоков будут границы предложений, в которых происходит переход от одного ключевого понятия к другому.

4 Методы установления смыслового тождества текстовых фрагментов

4.1 Модель представления текста

После деления текста на смысловые фрагменты можно приступить к сравнению текстов между собой, для этого необходимо представить их в формализованном виде. Мы решили выбрать модель представления текста, описанную в статьях [17-18], где смысловая структура текста была представлена в виде совокупности нормализованных наименований понятий и связей между ними. Такая смысловая структура текста была названа в этих работах его формализованным смысловым описанием.

В состав формализованного смыслового описания документа включены наименования понятий, сопровождаемые коэффициентом, определяющим степень их смысловой значимости в тексте. Этот весовой коэффициент будет зависеть от их синтаксической роли в предложении и частоты появления в тексте. Подробнее о критериях отнесения слов и словосочетаний к ФСОД изложено в работе [17].

В наших исследованиях мы будем использовать представленное в статьях [17-18] определение формализованного смыслового описания документа (ФСОД), под которым будем называть упорядоченное множество $F = \{Su_i \mid i \in [1, n_F]\}$, где

n_F - количество элементов в формализованном смысловом описании документа;

$Su_i = (Nc_i, w_i, R_i)$ - i-ый элемент ФСОД;

Nc_i — наименование понятия;

w_i - весовой коэффициент, соответствующий наименованию понятия;

R_i - множество связей, относящихся к данному элементу ФСОД.

Таблица 3

Проверка на заимствования документа с именем 0220xxxxx84

Название сравниваемого документа	Всего предложений	Совпало (с вероятностью > 90%)	% совпавших предложений (в тексте)	% совпавших предложений (в фрагментах)
0220xxxxx04	2193	75	8.1	78.6
0220xxxxx05	1650	43	4.6	86.3
0220xxxxx18	1179	97	10.5	88.2

4.2 Отождествление наименований понятий

Для того чтобы решить проблему правильного построения формализованного смыслового описания необходимо решить проблему отождествления наименований понятий. Наименования понятий в текстах могут быть представлены словами и словосочетаниями. При этом описание одинаковых понятий или ситуаций часто может выполняться в терминах различной степени общности и с помощью различных языковых средств. Например, в различных контекстных окружениях наименования понятий могут описываться с использованием явлений словоизменения и словообразования, а также явлений синонимии и гипонимии. Все эти явления существенно затрудняют распознавание и сравнение между собой текстовых форм представления наименований понятий. В связи с этим для отождествления различных форм представления наименований необходимо их приводить к канонической форме. Такое приведение может выполняться на различных уровнях обобщения их смысла: словоизменения, словообразования и синонимии.

Обобщение на уровне словоизменения производится путем последовательного приведения каждого слова к его канонической форме [3-4].

Обобщение на уровне словообразования производится путем выделения у опорных слов их словообразовательных основ и пословной нормализации определяющих их слов. При этом определяющие слова сортируются в их лексикографическом порядке.

Обобщение на уровне синонимии производится по словарю синонимичных слов и фразеологических словосочетаний объемом 450 тыс. словарных статей (Словарь составлен при участии авторов путем обработки двуязычных словарей общим объемом 4,5 млн. словарных статей). В этом словаре представлены смысловые инварианты слов или словосочетаний, один из которых выступает в качестве канонической формы представления смысла наименований понятий словарной статьи. Процедура сведения различных форм представления наименования понятий выполняется путем замены наименования понятия на его каноническую форму представления. Отождествление наименований производится путем сравнение полученных канонических форм двух словосочетаний. В случае успешного сравнения их текстовые формы считаются формами представления одного и того же понятия.

4.3 Процесс установления смыслового тождества текстовых фрагментов

После формирования формализованных смысловых описаний документов, можно перейти к задаче установления смыслового тождества текстовых фрагментов документа. Для этого требуется сопоставить полученные формализованные смысловые описания этих двух текстов. Поскольку фактически мы сопоставляем два графа, нами было решено использовать метод сравнения, который используется для поиска изоморфных пересечений двух графов [19]. Как пишут авторы, этот 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 пирамидой. 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.

5 Заключение

Описанные в первой части работы методы были реализованы, и полученное программное обеспечение позволило произвести исследования, целью которого являлось выявление общих элементов в научно-технических документах. Результатом данного исследования стало создание алгоритма выявления смысловых фрагментов текста с последующим построением их формализованных смысловых описаний и сравнения методом, представленным в работе [19]. Реализация данного метода будет большим шагом вперед в решении задачи установления заимствований в

структурированных научно-технических документах, ведь до сих пор в данной области семантические методы не получили широкого распространения в связи со сложностью их реализации. В данный момент идет работа по созданию полноценного программного комплекса, реализующего предложенные методы и включающего полный цикл решения задачи нахождения заимствований в текстах документов. Дальнейшим развитием данной системы станет использование данных методов для автоматического проведения экспертизы научно-технических документов на научную новизну, что позволит выявлять заимствования не только текста, но и смысла.

Литература

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

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

[3] Белоногов Г.Г., Калинин Ю.П., Хорошилов А.А. Компьютерная лингвистика и перспективные информационные технологии. Теория и практика построения систем автоматической обработки текстовой информации — М.: Русский мир, 2004. – 264 с.

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

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

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

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

[8] Чугреев В.Л. Модель структурного представления текстовой информации и метод ее тематического анализа на основе частотно-контекстной классификации. Диссертация на соискание ученой степени кандидат технических 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 Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL’2007, Переславль, Россия, 2007. – Том 1, С. 166-174.

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

[11] A. Broder, S. Glassman, M. Manasse and G. Zweig. Syntactic clustering of the Web. Proc. of the 6th International World Wide Web Conference, April 1997.

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

[13] Р.В. Шарапов, Е.В. Шарапова Система проверки текстов на заимствования из других источников // Труды 13 ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL’2011, Воронеж, Россия, 2011. – Том 1.

[14] 2. Авдеева Н.В., Ботов П.Ю., Букаев А.С., Вислый А.И., Груздев И.А., Житлухин Д.А., Романов М.Ю., Чехович Ю.В. Внедрение системы «Антиплагиат» в Российской государственной библиотеке // Материалы конференции «Интеллектуализация обработки информации» – октябрь, 2010. – С. 499–503.

[15] Никитов, А. В. Плагиат в работах студентов и аспирантов: проблема и методы противодействия / А. В. Никитов, О. А. Орчаков, Ю. В. Чехович // Университет. управление: практика и анализ. - 2012. - № 5. - С. 61 - 69.

[16] Антиплагиат [Электронный ресурс]. — Режим доступа: http://www.antiplagiat.ru

[17] 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, вып.8, 2012

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

[19] Агарков А.В. Метод сравнения двух графов за полиномиальное время. // Научно-теоретический журнал "Искусственный интеллект" №.4'2003.

[20] 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 октября 2009 г.

[21] Hartrumpf, Sven; Tim vor der Brück; and Christian Eichhorn (2010a). Detecting duplicates with shallow and parser-based methods. In Proceedings of the 6th International Conference on Natural Language Processing and Knowledge Engineering (NLPKE), pp. 142-149. Beijing, China

[22] Carmen Banea, Samer Hassan, Michael Mohler, and Rada Mihalcea, UNT: A Supervised Synergistic Approach to Semantic Text Similarity, Proceedings of the Sixth International Workshop on Semantic Evaluation SemEval 2012

Semantic methods for solving a problem of automatic detection of plagiarism in structured scientific and technical documents

Victor N. Zakharov, Alexey A. Khoroshilov

This paper presents the semantic methods for plagiarism detection in structured scientific and technical documents. The existing systems allow to detect plagiarism only when the plagiarism is made by copying text fragments without changes or with minor changes of its structure or lexical composition. The use of semantic methods for text analysis makes it possible to reveal the conceptual structure of the text and recognize more sophisticated cases of intended changes in the plagiarized texts, for example, to determine the cases of substitution of words or word combinations by their semantic invariants, changes in text splitting into sentences, replacement of the text fragments. This paper also grounds the need to detect text fragments, relating to description of structural document elements (for example, standard headings for all documents) and to exclude them from consideration.