Метод выявления имплицитно выраженных заимствований в научно-технических текстах на основе их концептуального анализа

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

Аннотация

В работе рассматривается процесс автоматического выявления неявно выраженных заимствований в текстах документов, основанный на сопоставлении их формализованных представлений. При решении данной задачи была разработана модель представления смысловой структуры текстов и методы формализации и установления смысловой близости между фрагментами сравниваемых текстов, а также методы выявления схожих по смысловой структуре фрагментов текстов. Основным преимуществом данного метода является то, что он позволяют эффективно выявить различного рода заимствования, включая самые сложные случаи – неявно выраженные заимствования.

1 Введение

1.1 Проблема плагиата и пути ее решения

В процессе информатизации общества у человека появилась возможность получать огромное количество информации по интересующей его проблеме, не прилагая для этого больших усилий. В связи с этим появились уникальные возможности для использования научного и творческого потенциала, накопленного человечеством, для проведения исследований и получения новых результатов интеллектуальной деятельности, что позволяет обеспечивать непрерывное развитие той или иной области знаний. Но в то же время эти возможности влекут за собой неизбежные проблемы, связанные с появлением возможности присвоения авторства на чужие результаты интеллектуальной деятельности. Многие недобросовестные авторы, стараясь скрыть факт заимствования текста или его фрагмента, изменяют структуру текста, например, используя синонимы слов и словосочетаний, добавляя или удаляя слова, разбивая или объединяя предложений. Имеющиеся же системы, предназначенные для поиска заимствований в текстах документов, способны выявить лишь факты прямого заимствования. Поиск таких заимствований производится, как правило, путем сопоставления элементов их текстовых представлений. Это связано с тем, что в процессе выявления заимствований эти системы не учитывают смысловую структуру текста, а рассматривают текст как последовательность слов. Между тем, текст – это не множество слов и их последовательностей, и при установлении смысловой близости документов нужно сопоставлять не текстовое представление, а его формализованное смысловое содержание.

1.2 Цели исследования

Целью исследований является решение проблемы выявления имплицитно выраженных заимствований в текстах документов. В соответствии с указанной целью в работе поставлены следующие задачи:

- Исследовать и разработать модели представления смыслового содержания текстов документов.
- Исследовать и разработать методы унификации смыслового представления наименований понятий (с учетом явлений словоизменения и словообразования, а также синонимии и гипонимии).
- Исследовать и разработать методы и алгоритмы автоматического установления смысловой близости и смысловой схожести сравниваемых документов.
- Разработать алгоритм процесса выявления заимствований в текстах документов.
- Разработать критерии выявления заимствований в текстах документов.
- Провести экспериментальное исследование, устанавливающее достоверность теоретических концепций и эффективность разработанных методов выявления заимствований в текстах документов.

1.3 Обзор существующих средств и методов выявления заимствований в текстах документов

В начале 90-х годов XX века для борьбы с недобросовестными авторами начали создаваться

Труды XVII Международной конференции DAMDID/RCDL'2015 «Аналитика и управление данными в областях с интенсивным использованием данных», Обнинск, 13-16 октября 2015

системы, предназначенные для поиска плагиата. Наиболее известными системами, используемыми за рубежом, являются TurnItIn , SafeAssign, CopyScape, WriteCheck, iThenticate, PlagAware, PlagScan, Copyscape, CheckForPlagiarism.net, PlagiarismDetection.org. С середины 2000-х годов такие системы начали появляться и для русского языка. Одной из наиболее известных систем является разработанная компанией Forecsys система Антиплагиат. Она используется во многих ВУЗах, академических структурах, а также государственных библиотеках. Среди менее известных программных продуктов можно отметить такие системы, как eTXT Антиплагиат, Advego Plagiatus и Text.ru. Одним из главных требований к таким системам является возможность доступа к обширной базе знаний, содержащей документы. Эти источники в тоже время могут служить в качестве возможных источников заимствований. Специализированное программное обеспечение для поиска плагиата часто использует либо собственную базу документов, либо использует пространство Интернета путем генерирования специальных запросов для глобальных поисковых систем. Эти системы также различаются применением различных подходов для сравнения текстов документов между собой. Рассмотрим наиболее часто встречающиеся модели представления текстов и методы их сравнения.

Векторная модель представления текста. Идея данной модели [21] заключается в представлении документа в виде вектора в n-мерном евклидовом пространстве, причем размерность документа определяется числом термов во всей коллекции документов. Каждому терму ставится в соответствие характеристика, определяющая его значимость, часто для этой цели используется частота появления данного терма в документе. При применении данной модели важно понимать, что авторы существенно упростили смысловую структуру документа, и текст представляется в виде набора слов, причем порядок их следования не учитывается. Для того чтобы выявить насколько близки по смыслу тексты необходимо найти меру близости двух векторов, которые соответствуют текстам.

Методы, основанные на вычислении сигнатур. Основной идеей таких методов [22] является вычисление «сигнатуры» - числового значения соответствующего тексту документа. Соответственно, если эти сигнатуры совпадают, то документы считаются похожими. Один из наиболее известных таких методов, I-Match. Для определения похожих документов сначала составляется словарь термов, входящих в исходный документ, после этого определяется общая часть словарей составленных по документу и по корпусу текстов. Затем вычисляется I-Match сигнатура документа. Для этого словарь упорядочивается, а затем применяется хэш-функция. Полученная численная характеристика описывает исходный документ и позволяет эффективно сравнивать документы между собой.

Метод шинглов. Основанная идея данного метода заключается в представлении текста в виде множества последовательностей слов фиксированной длины – шинглов [12, 24, 25]. Эти последовательности должны состоять из соседних слов в порядке их следования, причем эти последовательности должны идти внахлест. После разбиения текста на такие последовательности для них считаются хэш-коды. Далее для сравнения документов необходимо выявить насколько совпадают множества хэш-кодов шинглов.

Семантические методы. Методы сравнения документов, основывающиеся на семантических методах обработки текстовой информации, имеют ряд преимуществ. Они позволяют сравнивать не цепочки слов, а смысловую структуру текста и, поэтому используя такие модели, можно выявлять не только простейшие случаи заимствований в текстах, но и более сложные, когда автор целенаправленно меняет текст, если при этом сохраняются взаимосвязи между понятиями в тексте. В качестве инструмента для решения таких задач, применяются процедуры морфологического и семантико-синтаксического анализа. Однако можно выявить и недостатки такого подхода. Основная модель для представления текста в существующих работах [2,14,15,23,26] – концептуальный граф или ему подобная структура, построение которого достаточно трудоемкая задача [13], требующая наличия сложного семантического инструментария.

Все описанные в данном разделе методы имеют ряд недостатков, например, в первых трех - различные операции, выполняемые в процессе сопоставления их текстового представления, производятся в отрыве от анализа смысловой составляющей этих текстов и эти методы позволяют достаточно успешно выявлять в основном только эксплицитно выраженные заимствования. Более сложные семантико-статистические и глубинные семантические методы [19-20] позволяют решать частичные задачи установления имплицитно выращенных заимствований. Но они, вследствие их ориентации на сложные семантические инструменты (такие как семантические словари, тезаурусы или онтологии), также не получили широкого распространения, обусловленного недостаточной полнотой покрытия лексики текстов этими инструментами. На наш взгляд для решения проблемы эффективного выявления имплицитно выраженных заимствований необходимо разработать методы и средства формализации смысловой структуры текстов, выявления близких по смыслу фрагментов текстов и установления их смысловой схожести между собой [8].

2 Выявления имплицитно выраженных заимствований в текстах

2.1 Теоретическое представление о смысловой структуре текста

При моделировании понимания смыслового содержания текстов важно исходить из правильных представлений о его смысловой структуре. По современным представлениям наиболее информативными и наиболее устойчивыми единицами смысла являются понятия [5]. Они занимают центральное место в языке и речи, с их помощью описывается смысловое содержание текстов и именно они являются теми базовыми строительными блоками, на основе которых формируются смысловые единицы описания текста.

Рассматривая проблему единиц смысла языка и речи, нельзя хотя бы поверхностно, не коснуться вопроса о внутренней структуре понятий, представляющих план содержания этих единиц. И причина здесь состоит в том, что, выражаясь словами основоположника науки лингвистики – Фердинанда де Соссюра [5], конкретные языковые сущности не даны нам в непосредственном наблюдении. “Смысл” же понятия проявляется в полной мере только через всю систему его отношений со всеми другими понятиями языка [9,10,11,17]. По сути дела он так же неисчерпаем, как неисчерпаем язык в целом и как неисчерпаемо человеческое мышление [5]. Базируясь на этом утверждении можно ввести такие понятия, как понятия глобальной и локальной связности текстов. При этом понимается, что глобальная связность обеспечивает раскрытие содержания текста. Локальная связность обеспечивает раскрытие смысла понятия на основе его контекста. Под смысловой связанностью текста или его фрагмента будем понимать совокупность взаимосвязанных наименований понятий, расположенных в тексте в определённом порядке и отражающих основное смысловое содержание текста или его фрагмента.

Преобразование текстового представления в его формализованное смысловое представление обеспечивает сопоставление текстов по их смысловому содержанию[1,3,4,16]. Результатом такого сопоставления является установление смысловой близости документов. Под смысловой близостью мы понимаем, что в двух текстах или их фрагментах, описываемых одной и той же совокупностью наименованиями понятий, некие ситуации, могут быть близкими по смыслу , но не обязательно идентичными. Идентичными могут быть ситуации, если они описывается одинаковыми наименованиями понятий и имеют схожий (не обязательно одинаковый) порядок их следования, удовлетворяющий условиям локальной и глобальной схожести. Условием локального смыслового сходства является сходство контекстного окружения идентичных наименований понятий в двух текстах или их фрагментах. Условием глобального смыслового сходства является сходство порядка следования наименований понятий в сравниваемых двух текстах или их фрагментах.

Здесь необходимо пояснить, что поскольку смысл наименования понятия в значительной степени определяется его контекстным окружением, а нашей модели смысл текста определяется как совокупность взаимосвязанных наименований понятий, то локальная схожесть текста будет определяться каждым конкретным наименованием понятия и его контекстным окружением (упорядоченной последовательность наименований понятий слева и справа от данного понятия.) Аналогично под глобальной схожестью смыслового представления фрагментов текстов понимается совокупность взаимосвязанных наименований понятий, расположенных в тексте в определённом порядке, в котором каждое понятие удовлетворяет условию локальной схожести.

2.2 Модель процесса выявления имплицитно выраженных заимствований в текстах

В качестве модели для представления смыслового содержания текста будем использовать совокупность выявленных в тексте наименований понятий - концептуальный образ документа [6,7], дополненный их контекстным окружением наименований понятий (КОДКО).

Таким образом КОДКО $= \{НП_i, K_i | i \in [1, n_{НП}]\}$, где $НП_i = (t_i, Адр_i)$ - i-ое наименование понятия; $n_{НП}$ - количество наименований понятий;

t_i - текстовое представление i – ого наименования понятия;

$Адр_i$ - адрес вхождений i – ого наименования понятия в тексте;

K_i - множество контекстов i – ого наименования понятия, где эти контексты описываются аналогичным образом

$К_i = \{НПК_i | i \in [1, n_{НПК_i}]\}$, $НПК_i = (t_i, Адр_i)$.

Для хранения результатов установления локального смыслового сходства документов и подсчета коэффициентов смысловой близости между фрагментами текстов будет использоваться матрица

$$МЭ = \begin{pmatrix} м_{11} & м_{12} & .. & м_{1n_{Ф1}} \\ м_{21} & м_{22} & .. & м_{2n_{Ф1}} \\ : & : & м_{ij} & : \\ м_{n_{Ф2}1} & м_{n_{Ф2}2} & .. & м_{n_{Ф2}n_{Ф1}} \end{pmatrix}$$

размерности $n_{Ф1} \times n_{Ф2}$, где $n_{Ф1}$ - количество элементов формализованного смыслового описания 1-ого из сравниваемых документов, $n_{Ф2}$ - количество элементов формализованного смыслового описания 2-ого из сравниваемых документов,

Где $м_{ij}$ - численная характеристика выполнения условия локального смыслового сходства. В случае $м_{ij} = 0$ данное условие – не выполнено, при $м_{ij} > 0$ - выполнено частично, а при $м_{ij} = 0$ - выполнено полностью.

Если $снп(НП_{pi}, НП_{qj}) = 0$, то $м_{ij} = 0$, иначе

$$м_{ij} = \frac{снп(НП_{pi}, НП_{qj})}{3} + \frac{2(\sum_{l,m=0}^{n_{НПК_{pi}}, n_{НПК_{qj}}} снп(К_{pil}, К_{qjm}))}{3(n_{НПК_{pi}} + n_{НПК_{qj}})}$$

,где $снп(НП_{pi}, НП_{qj})$ - функция определения эквивалентности словосочетаний, причем $снп(НП_{pi}, НП_{qj}) \in [0,1]$, $НП_{pi}$ - i-ый элемент формализованного смыслового описания p-ого документа, $НП_{qj}$ - j-ый элемент формализованного смыслового описания q-ого документа.

Поскольку функция $снп(НП_{pi}, НП_{qj})$ - возвращает значения отличные от 0 значительно меньшее количество раз, чем нулевые значения, матрица S будет считаться разреженной. Соответственно данную матрицу можно заменить 3-я векторами.

$$ВЗ = \begin{pmatrix} вз_1 \\ : \\ : \\ вз_{n_{вз}} \end{pmatrix}$$ – вектор значений из матрицы МЭ,

для которых выполняется условие $м_{ij} > 0$, причем $i \in [0, n_{Ф2}]$, $j \in [0, n_{Ф1}]$.

$$ВК_p = \begin{pmatrix} вк_{p1} \\ : \\ : \\ вк_{pn_{вз}} \end{pmatrix}$$ – вектор индексов из формализованного смыслового описания p-ого документа, соответствующих значениям из вектора ВЗ.

$$ВК_q = \begin{pmatrix} вк_{q1} \\ : \\ : \\ вк_{qn_{вз}} \end{pmatrix}$$ – вектор индексов из формализованного смыслового описания q-ого документа, соответствующих значениям из вектора ВЗ.

Условием глобального смыслового сходства является сходство порядка следования наименований понятий, но поскольку порядок следования наименований понятий учтен при подсчете коэффициентов $м_{ij}$,с точностью до перестановок, которые возможны в тексте из-за особенностей естественного языка. После этого производится поиск последовательностей наименований понятий, у которых значения локальной смысловой схожести $вз_i$ выше некого заданного порога, в исследованиях использовались значения $вз_i > 0.65$. Затем для этих последовательностей вычисляются меры выполнения условия глобального смыслового сходства. Данная задача сводится к вычислению среднего значения характеристик выполнения условия локального смыслового сходства, содержащихся в этих последовательностях наименований понятий. Эта величина и будет являться коэффициентом смысловой близости фрагментов текстов:

$$K_{сх} = \frac{\sum_{i=0}^{n_{взЦеп}} вз_{Цеп_i}}{n_{взЦеп}}$$

Где $вз_{Цеп}$ - элемент вектора ВЗ, принадлежащий найденной цепочке, $n_{ВЗ}$ - число элементов в цепочке.

2.3 Алгоритм выявления имплицитно выраженных заимствований в текстах

В результате проведенных исследований был разработан алгоритм выявления имплицитно выраженных заимствований в текстах документов. Необходимым условием для реализации этого алгоритма является использование эталонного концептуального словаря (словаря ЭКС), обеспечивающего покрытие понятийного состава не менее чем на 90%. Словарь ЭКС (объемом 1.7 млн. наименований понятий), включающий понятийный состав широкого спектра тематик, был ранее создан в рамках проекта МетаФраз [7] путем автоматизированного анализа 30 млн. рефератов документов, содержащихся в базах данных ВИНИТИ [3] и использовался в данном эксперименте для выявления наименований понятий в текстах. Для решения задачи приведения слов и словосочетаний к их унифицированным смысловым представлениям производится ряд трансформаций: 1) все слова этого понятия приводятся к канонической форме; 2) производится перестановка слов (главное слово на первой позиции, зависимые слова в лексико-грамматическом порядке); 3)если полученное формализованное представление

совпадает с одним из входов словаря фразеологических синонимов (объёмом 500 тыс. словарных статей) производится его замена на унифицированное формализованное представление (замена на доминантный синоним).

В результате обработки текста процедурами семантико-синтаксического и концептуального анализа в тексте по словарю ЭКС выявляются и формализуются наименования понятий с указанием номера предложения и места в нем. Порядок их следования в тексте с помощью указанных адресов всегда сохраняется. Контекстное окружение каждого понятия задается граничными условиями и определятся упорядоченной (по тексту) совокупностью понятий слева и справа от него. (Например, 7 понятий слева и 7 понятий справа).

Общий порядок выполнения алгоритма выявления имплицитно выраженных заимствований включает следующие этапы:

Шаг 1. В анализируемом тексте с помощью эталонного концептуального словаря выявляется совокупность значимых наименований понятий с указанием местоположений этих понятий в тексте.

Шаг 1. Каждое наименование понятия с помощью процедуры автоматической пословной нормализации и словаря унифицированного формализованного представления наименований понятий приводится к унифицированной форме.

Шаг 2. Производится поиск совпадающих наименований понятий в массиве формализованных представлений документов.

Шаг 3. Для рассматриваемого документа устанавливаются перечень документов близких ему по смысловому содержанию.

Шаг 4. Для пары документов - рассматриваемого документа и каждого из документов, найденных в п. 4 устанавливаются пары наиболее близких по смысловому содержанию фрагментов анализируемых текстов.

Шаг 5. Для каждой установленной в п.5 пары близких по смыслу фрагментов текстов определяются локальная смысловая схожесть всех наименований понятий этих фрагментов.

Шаг 6. Выбираются последовательности наименований, имеющих значения локальной смысловой схожести выше заданного порога. Для каждой такой последовательности наименований понятий обоих текстов вычисляется степень их глобальной смысловой схожести.

3 Эксперимент по выявлению имплицитно выраженных заимствований в текстах

Таблица 1. Средняя полнота и точность выявления заимствований в текстах документов

Полнота (разработанный метод)	Полнота (метод шинглов)	Точность (разработанный метод)	Точность (метод шинглов)
0.67	0.48	0.94	0.96

В процессе кластеризации при установлении степени смысловой близости документов массива были заданы пороговые критерии: степень смысловой близости документов (более 12%) и минимальное количество совпавших наименований понятий в сравниваемых документах (более 30 элементов). На основе этих критериев было получено 17 кластеров документов, в которых количество документов олебалась от 23 до 117. Далее для каждого кластера были посчитаны численные значения полноты и точности поиска заимствований для результатов, полученных обоими способами. Усредненные результаты проведенного исследования приведены в таблице 1.

4 Заключение

Разработанный подход к решению проблемы выявления имплицитно выраженных заимствований в научно-технических текстах на основе их концептуального анализа на этапе предварительных исследований показал свою высокую эффективность при установлении различного рода заимствований, включая самые сложные случаи – имплицитно выраженные заимствования. Этот подход базируется на следующих разработанных автором методах формализации и анализа смысловой структуры текстов:

- Метод автоматического выявления наименований понятий в текстах документов.
- Метод автоматического установления смысловых связей между наименованиями понятий.
- Метод автоматического приведения текстового представления понятий к их унифицированному формализованному представлению.
- Метод автоматического разбиения анализируемых текстов на взаимосвязанные фрагменты.
- Метод установления смысловой схожести взаимосвязанных фрагментов анализируемых текстов.

В заключение необходимо отметить, что

проблема эффективного выявления неявно выраженных заимствований требует комплексного подхода к ее решению, связанного с тем, что наряду с необходимостью разработки методов и средств формализации смысловой структуры текстов, требуется также создание и постоянная актуализация декларативных средств для анализа текстовой информации.

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

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

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

[6] 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.

[7] Захаров В.Н., Хорошилов А.А. Автоматическая оценка подобия тематического содержания 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.

[8] 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.

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

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

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

[12] Зеленков Ю.Г., Сегалович И.В. Сравнительный анализ методов определения нечетких дубликатов для WEB-документов // Труды 9-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» RCDL’2007: сб. работ участников конкурса / Переславль-Залесский, Россия, 2007.

[13] Богатырев М.Ю. , Латов В.Е. , Столбовская И.А. Применение концептуальных графов в системах поддержки электронных библиотек // Тр. 9-й Всеросс. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL 2007. — Переславль-Залесский, Россия, 2007. — Т. 2. — С. 104—110.

[14] А. В. Палагин, С. Л. Кривой, Н. Г. Петренко Концептуальные графы и семантические сети в системах обработки естественно-языковой информации // Математичні машини і системи. -Київ, 2009,N N 3.-С.67-79

[15] 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.

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

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

[18] 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.

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

[20] 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.

[2[1] Salton, G.; Wong, A.; Yang, C. S. (1975). "A vector space model for automatic indexing" / Communications of the ACM Volume 18 Issue 11, New York, NY, USA, Nov. 1975 Pages 613-620., Salton et al. 1994.

[22] Abdur Chowdhury, Ophir Frieder, David Grossman, Mary Catherine McCabe // Collection statistics for fast duplicate document detection // Journal ACM Transactions on Information Systems (TOIS) TOIS Homepage archive Volume 20 Issue 2, April 2002, Pages 171-191.

[23] Vor der Brück T., Hartrumpf S. A readability checker based on deep semantic indicators// In

Human Language Technology. Challenges of the Information Society – 2009. – V. 5603 of Lecture Notes in Computer Science (LNCS). – P. 232-244. Berlin, Germany: Springer.

[24] A. Broder. On the resemblance and containment of documents. Compression and Complexity of Sequences (SEQUENCES'97), pages 21-29. IEEE Computer Society, 1998.

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

[26] 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.

A Method for Detecting Implicit Plagiarism in Scientific and Technical Texts on the Basis of Their Conceptual Analysis

Alexey A. Khoroshilov

The paper presents the process of automatic plagiarism detection in documents on the base of comparison of their formalized representations. In solving this problem we developed a model of the semantic structure of texts. To detect plagiarism, we developed an algorithm for detection of similar semantic fragments and a method for identification of semantic similarity between text fragments. The main advantage of this method is that it makes it possible to detect not only minor changes in the structure or lexical structure of the text, but also more complicated cases of intended changes in the plagiarized texts.