В.Н. Захаров, А.А. Хорошилов, А.А. Хорошилов

Метод автоматического выявления неявно выраженных заимствований в научно-технических текстах[1]

Аннотация. В работе рассматривается процесс автоматического выявления неявно выраженных заимствований в текстах документов, основанный на сопоставлении их формализованных представлений и вычислении мер локальной смысловой схожести понятий и глобальной смысловой схожести фрагментов текстов. 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 позволяет эффективно выявить различного рода заимствования, включая самые сложные случаи – неявно выраженные заимствования. В ходе исследования результаты работы метода были сопоставлены с результатами, полученными с применением метода «шинглов».

Ключевые слова: Ключевые слова: выявление заимствований, автоматизированная обработка текстов, формализованное описание текста, смысловая структура, лингвистическое программное обеспечение, декларативные средства.

Введение

Наличие заимствований в работах недобросовестных авторов является серьезной проблемой, значительно влияющей на качество образования и науки в стране. В настоящее время в зарубежной академической практике западных университетов и научных журналов существуют документы, регулирующие правила заимствований текста и оформления соответствующих ссылок на источники, а также четко прописаны критерии отнесения некорректных заимствований к плагиату в различных формах. Плагиатом, как правило, считается любое использование чужих идей и высказываний без должной отсылки к источнику. Заимствованием также считается неадекватный пересказ текста другого источника, при котором изложение фрагмента осуществляется путем замены некоторых слов в исходном тексте с сохранением его структуры. Даже если при этом дается полная ссылка на источник, и адекватный пересказ, но не сопровождающийся указанием на источник заимствования идей.

В нашей стране, к сожалению, критерии выявления плагиата регламентированы не столь серьезно. Несмотря на это во многих ведущих вузах страны введены положения, которые подробно определяют ответственность учащихся за любые виды заимствований в своих работах. Так, например, в положении “О порядке применения дисциплинарных взысканий при нарушениях академических норм в написании письменных учебных работ в «Государственном университете – Высшей школе экономики» плагиат определяется как использование в письменной работе чужого текста, опубликованного в бумажном или электронном виде, без полной ссылки на источник или со ссылками, но когда объем и характер заимствований ставят под сомнение самостоятельность выполненной работы или одного из ее основных разделов. Плагиат может осуществляться в двух

[1] Работа выполнена при поддержке Министерства образования и науки РФ, уникальный идентификатор проекта RFMEFI60414X0139.

видах: дословное изложение чужого текста; парафраза – изложение чужого текста с заменой слов и выражений без изменения содержания заимствованного текста.

Если первый вид заимствований выявляют большинство систем, анализирующих текст на наличие плагиата, то выявление второго вида заимствований является довольно сложной задачей. В этой статье авторами дается описание метода, который позволяет решить эту непростую задачу.

1. Обзор существующих моделей и методов

В настоящее время для решения задачи поиска плагиата используется большое число моделей представления текстов и методов их сравнения. Рассмотрим наиболее часто встречающиеся из них.

1. Векторная модель представления текста. Идея данной модели [1] 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.

2. Методы, основанные на вычислении сигнатур. Основной идеей таких методов [2] 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.

3. Метод шинглов. Основная идея данного метода заключается в представлении текста в виде множества последовательностей слов фиксированной длины – шинглов [3-5]. Эти последовательности должны состоять из соседних слов в порядке их следования, причем 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.

4. 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 [6, 7] – концептуальный граф или ему подобная структура, построение которого – достаточно трудоемкая задача [7], требующая наличия сложного семантического инструментария.

Все описанные 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. Более сложные семантико-статистические и глубинные семантические методы [8, 9] ориентированы на инструменты (такие как семантические словари, тезаурусы или онтологии), которые довольно непросты в реализации. Поэтому, на наш взгляд, для решения проблемы эффективного выявления семантической близости содержания текстов, необходимо разработать методы и средства формализации смысловой структуры текстов, выявления близких по смыслу фрагментов текстов и установления их смысловой схожести между собой [10].

2. Модель процесса выявления неявно выраженных заимствований в текстах документов

В качестве базовой теоретической концепции авторами использовалась концепция проф. Г.Г. Белоногова и проф. Р.С. Гиляревского, констатирующая, что смысловое содержание текстов выражается с помощью единиц смысла, входящих в их состав. По их мнению, наиболее устойчивыми единицами смысла являются понятия. Г.Г. Белоногов определяет термин «понятие» как «социально значимый мыслительный образ, за которым в языке закреплено его наименование в виде отдельного слова или, значительно чаще, в виде устойчивого фразеологического словосочетания…» [11].

Понятия 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.

В работе [12] вводятся понятия глобальной и локальной связности текстов. При этом констатируется, что глобальная связность обеспечивает раскрытие темы документа, а локальная связность проявляется во взаимосвязи между соседними единицами текста. В соответствии с нашей моделью под глобальной смысловой связностью текста или его фрагмента будем понимать смысловую связь совокупности наименований понятий текста или его фрагмента, расположенных в определенном порядке. Под локальной смысловой связностью текста или его фрагмента будем понимать смысловую связь конкретного наименования понятия и его контекстного окружения.

Преобразование текстового представления в его формализованное смысловое представление дает возможность сопоставления текстов по их смысловому содержанию. Такое сопоставление смыслового содержания текстов, обеспечивающее выявление близких по смыслу фрагментов текстов, на наш взгляд, должно удовлетворять следующим условиям:

- В двух текстах должна быть пересекающаяся совокупность наименований понятий. Их число должно быть равно или превышать наименования понятий, входящих в состав единичного высказывания [13].
- В двух таких текстах должны быть фрагменты, в которых концентрация пересекающихся наименований понятий превышает пороговое значение. Эти фрагменты должны иметь соизмеримые размеры.
- Фрагменты текстов должны быть сходными по составу наименований понятий и порядку их следования.

Определение схожего порядка следования наименований понятий в тексте или его фрагменте базируется на предположении, что смысл наименований понятий в значительной степени определяется их контекстным окружением [14]. В нашей модели смысл текста определяется как смысловое содержание совокупности взаимосвязанных наименований понятий, расположенных в нем в определенном порядке. Идентичные по смыслу тексты или их фрагменты должны удовлетворять условиям локальной и глобальной смысловой схожести. Локальная смысловая схожесть (ЛСС) наименований понятий текста определяется как сходство контекстного окружения идентичных наименований понятий в двух текстах или их фрагментах. Глобальная смысловая схожесть (ГСС) текстов или их фрагментов определяется как сходство состава идентичных наименований понятий и порядка их следования в текстах или их фрагментах. Каждое понятие этого фрагмента также должно удовлетворять условию локальной смысловой схожести.

Предлагаемая модель позволяет выявить близкие по тематике тексты или их фрагменты, а затем определять среди них явно и неявно выраженные заимствования. Рассмотрим подробнее данную модель.

Документы, близкие по смыслу документу $Д_р$, будут выявляться в массиве МД , состоя-

щем из $n_{МД}$ документов. Для начала необходимо определить подмножество документов содержащих аналогичные наименования понятий. Для этого предварительно каждому документу массива, а также при обработке анализируемому документу, ставятся в соответствие их формализованные смысловые описания. Эти описания представляют собой совокупность номеров наименований понятий, выявленных в тексте, идентифицированных по словарю унифицированных формализованных представлений наименований понятий (УФПНП) и сопровождаемых адресами их вхождения в текст, которые представлены идентификатором текста, номером предложения в этом тексте и позицией наименования понятия в предложении.

Такое формализованное описание документа будем называть концептуальным образом документа (КОДом):

$$КОД = \{НП_i \mid i \in [1, n_{НП}]\};$$

$n_{НП}$ – количество элементов в концептуальном образе документа;

$$НП_i = (ННПС_i, Адр_i);$$

$НП_i$ – информация об i-ом наименовании понятия;

$ННПС_i$ – номер наименования понятия в словаре УФПНП;

$$Адр_i = \{ИТ_{ij}, ИП_{ij}, ПСП_{ij} \mid j \in [1, n_{Адр_i}]\};$$

$Адр_i$ – информация о местоположении наименования понятия в тексте;

$ИТ_{ij}$ – идентификатор текста, в котором находится наименование понятия;

$ИП_{ij}$ – идентификатор предложения, в котором находится наименование понятия;

$ПСП_{ij}$ – позиция наименования понятия в предложении.

Рассмотрим анализируемый документ $Д_p$ и массив имеющихся в хранилище документов МД:

$$МД=\{Д_j \mid j \in [1, n_{МД}]\};$$

$Д_j$ – j-й документ массива;

$n_{МД}$ – количество документов в массиве.

Каждый документ $Д_j$ массива предварительно обработан при помощи функции получения КОДа: $КОД_{Дj} = код(Д_j)$. В результате такому документу соответствует $КОД_j = \{НП_{ji} \mid i \in [1, n_{НП_j}]\}$. $код(Д)$ – функция получения кода, аргументом которой является текст документа. Аналогичным образом обрабатывается и анализируемый документ $Д_p$, которому соответствует $КОД_{Дp} = код(Д_p)$.

Для ограничения числа рассматриваемых документов массива необходимо выявить только те документы, в которых содержатся аналогичные наименования понятий. Для этого нужно сопоставить полученный $КОД_{Дp}$ анализируемого документа с КОДами всех документов массива. Процесс сопоставления выполняется путем попарного пересечения двух КОДов - анализируемого документа и документа массива. Найденные элементы КОДа ($КОД_{Д_pД_j}$) должны включать наименования понятий, входящие, как в документ $Д_p$, так и в документ $Д_j$, а адреса наименований понятий записываются из обоих текстов.

На Рис. 1 проиллюстрирован вышеописанный процесс сопоставления КОДов – анализируемого документа и документов массива. После этого формируется массив документов $МД' \in МД$, состоящий только из тех документов $Д_j$, для которых выполняется следующее условие:

$$n_{КОД_{Д_pД_j}} > k_{мспп},$$

где $n_{КОД_{Д_pД_j}}$ – размерность кода $КОД_{Д_pД_j}$, полученного при пересечении $КОД_{Дp}$ и $КОД_{Дj}$;

$k_{мспп}$ – коэффициент, соответствующий необходимому минимальному числу пересекающихся понятий в текстах документов.

На втором этапе процесса необходимо в каждом документе $Д_j \in МД'$, выявленного подмножества документов, определить фрагменты текстов, близкие по смыслу. Предварительно фрагменты текста автоматически выделяются в документе путем его разбиения на всевозможные отрезки текста различной длины, состоящие из контактно расположенных предложений, минимальный размер, которых может быть равен одному предложению, а максимальный – не превышает пороговую величину $k_{\max ф}$.

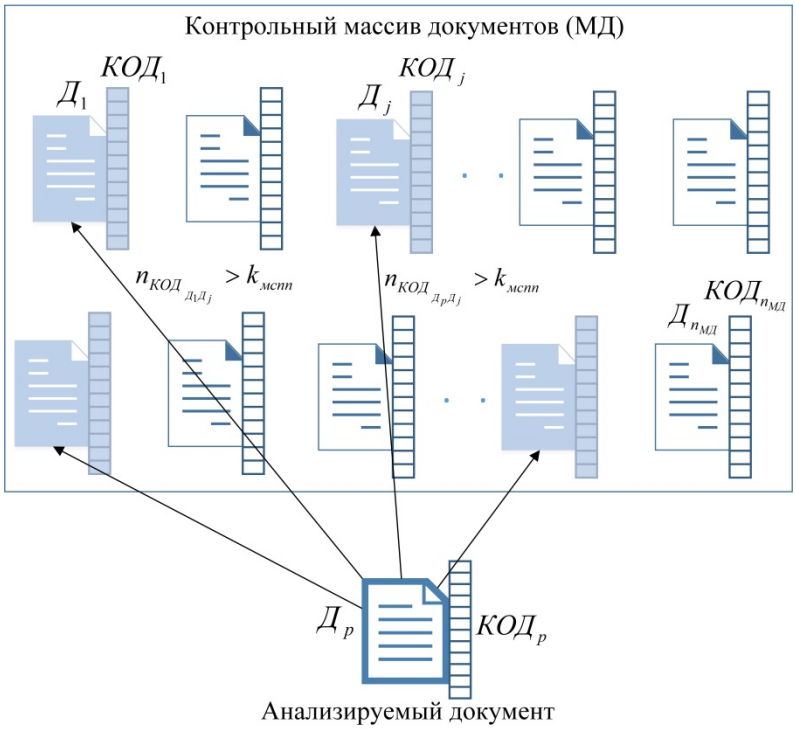

Рис.1. Сопоставление КОДов анализируемого документа и документов массива

Возьмем один из фрагментов текста $Д_p$ длиной в t предложений и начинающийся с l-го предложения и один из фрагментов текста $Д_j$ длиной в y предложений и начинающийся с z-го предложения. Этим фрагментам будут соответствовать списки информации о наименованиях понятий фрагментов (СИНПФ) $СИНПФ_{П_{p\,l}П_{p\,l+t}}$ и $СИНПФ_{П_{j\,z}П_{j\,z+y}}$ соответственно:

$$СИНПФ_{П_{p\,l}П_{p\,l+t}} = \{НП_{p\,l\,1},..,НП_{p\,l\,n_l},..,НП_{p\,l+t\,n_{l+t}}\};$$
$$СИНПФ_{П_{j\,z}П_{j\,z+y}} = \{НП_{j\,z\,1},..,НП_{j\,z\,n_z},..,НП_{j\,z+y\,n_{z+y}}\}.$$

Затем производится проверка условия сопоставимости фрагментов текстов:

$$\max(\text{len}(ФТ_{p\,l\,t}), \text{len}(ФТ_{j\,z\,y})) <$$
$$k_{мдо}\ \min(\text{len}(ФТ_{p\,l\,t}), \text{len}(ФТ_{j\,z\,y})),$$

где $ФТ_{p\,l\,t}$ – фрагмент текста $Д_p$;

$ФТ_{j\,z\,y}$ – фрагмент текста $Д_j$;

$k_{мдо}$ – коэффициент, определяющий, во сколько раз могут отличаться размеры сравниваемых фрагментов текстов;

len() – функция определения длины фрагмента текста (количество понятий);

min(), max() – функции выбора максимального и минимального значений соответственно.

Далее осуществляется сопоставление совокупностей наименований понятий, соответствующих этим фрагментам текстов. Для этого воспользуемся мерой Сёренсона. Мера смысловой близости этих двух фрагментов будет определяться по следующей формуле:

$$k_{сб} = \frac{2N(СИНПФ_{П_{p\,l}П_{p\,l+t}} \cap СИНПФ_{П_{j\,z}П_{j\,z+y}})}{N(СИНПФ_{П_{p\,l}П_{p\,l+t}}) + N(СИНПФ_{П_{j\,z}П_{j\,z+y}})},$$

где $N(СИНПФ_{П_{p\,l}П_{p\,l+t}})$ – количество понятий во фрагменте $П_{p\,l} \cup П_{p\,l+1} \cup .. \cup П_{p\,l+t}$;

$N(СИНПФ_{П_{j\,z}П_{j\,z+y}})$–количество понятий во фрагменте $П_{j\,z} \cup П_{j\,z+1} \cup .. \cup П_{j\,z+y}$;

$N(СИНПФ_{П_{p\,l}П_{p\,l+t}} \cap СИНПФ_{П_{j\,z}П_{j\,z+y}})$ – количество общих понятий для обоих фрагментов.

Аналогичным образом производится попарное сравнение всех фрагментов текстов, после этого – выбор текстовых фрагментов максимальной длины с коэффициентом смысловой близости больше порогового значения $k_{псб}$.

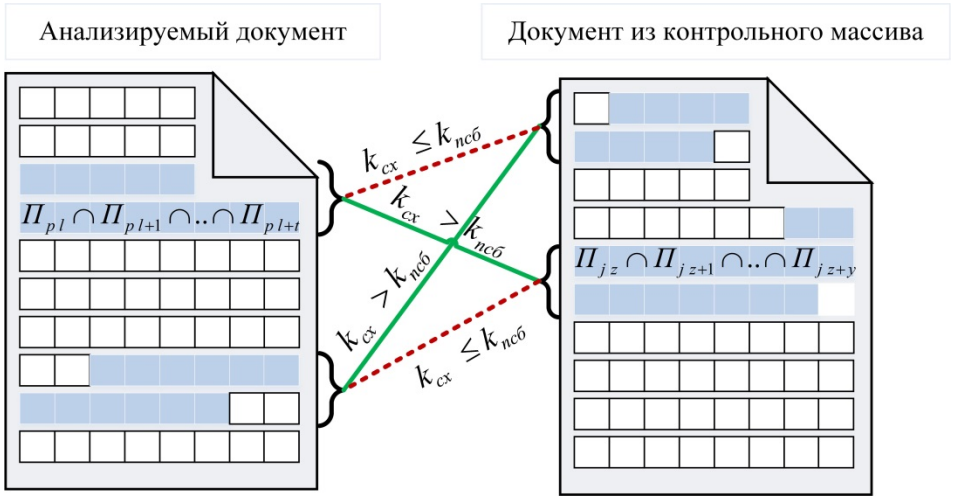

Рис. 2. Сопоставление фрагментов текстов анализируемого документа и документа массива

На Рис. 2 проиллюстрирован вышеописанный процесс сопоставления фрагментов текстов анализируемого документа и документа массива. Далее, если требуется, происходит проверка на смысловую идентичность, для этого необходимо установить смысловую схожесть фрагментов двух текстов, мера смысловой близости которых превышает пороговое значение. Для этого необходимо определить локальную смысловую схожесть наименований понятий, входящих в состав этих фрагментов, путем сопоставления окружающих его справа и слева понятий. В используемой модели для этого необходимо вычислить меру сходства элементов формализованного описания текстовых фрагментов, которые включают в себя контекстное окружение наименования понятия. После этого вычисляется мера глобального смыслового сходства, которая зависит от мер локального смыслового сходства.

На третьем этапе используется усовершенствованная модель представления смыслового содержания текста, где, в отличие от обычного КОДа, информация о наименованиях понятий дополняется контекстом. Такую модель будем называть концептуальным образом документа, дополненным контекстным окружением (КОДКО):

$$КОДКО = \{НП_i, K_i \mid i \in [1, n_{НП}]\},$$

где $НП_i = (ННПС_i, Адр_i, \text{ОСРНП}_i)$;

$НП_i$ – информация об i-м наименовании понятия;

$ННПС_i$ – номер наименования понятия в словаре УФПНП;

$Адр_i$ – адреса вхождений наименования понятия в тексте;

$n_{НП}$ – количество наименований понятий;

ОСРНП_i – символ обобщенной синтаксической роли i-го наименования понятия;

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

$$K_i = \{НПК_{ik} \mid k \in [1, n_{НПК_i}]\},$$

где $НПК_{ik} = (ННПС_{ik}, Адр_{ik}, \text{ОСРНП}_{ik}, \text{КЗК}_{ik})$;

$$\text{КЗК}_{ik} = \begin{cases} 1 & \text{, если в разных предложениях с НП}_i \\ 2 & \text{, если в одном предложении с НП}_i \end{cases};$$

КЗК_{ik} – коэффициент значимости контекста.

Затем вычислим значение меры $м_{ik}$ выполнения условия локального смыслового сходства для каждого наименования понятия из КОДКО сравниваемых документов. В случае $м_{ik} = 0$ данное условие – не выполнено, при $м_{ik} > 0$ – выполнено частично, а при $м_{ik} = 1$ – выполнено полностью.

Если $\text{снп}(\text{НП}_{pi}, \text{НП}_{jk}) = 0$, то $м_{ik} = 0$, иначе

$$м_{ik} = \frac{\text{снп}(\text{НП}_{pi}, \text{НП}_{jk})}{3} + \frac{2\text{ско}(\text{К}_{pil}, \text{К}_{jk\text{m}})}{3},$$

где ско() – функция сравнения контекстного окружения наименований понятий;

$$\text{ско}(\text{К}_\text{a}, \text{К}_\text{b}) = \begin{cases} 1 & \text{, фвзбк}(\text{К}_\text{a}, \text{К}_\text{b}) > 1 \\ \text{фвзбк}(\text{К}_\text{a}, \text{К}_\text{b}) & \text{, фвзбк}(\text{К}_\text{a}, \text{К}_\text{b}) < 1 \end{cases};$$

фвзбк() – функция вычисления значения близости контекстов;

$$\text{фвзбк}(\text{К}_\text{a}, \text{К}_\text{b}) = \frac{\sum_{c=0}^{n_{НПК_a}} \sum_{d=0}^{n_{НПК_b}} \text{фвппэ}(НПК_{ac}, НПК_{bd})}{4k_{мспп}};$$

$$\text{фвппэ}(НПК_{ac}, НПК_{bd}) = \begin{cases} 0 & , ННПС_{ac} \neq ННПС_{bd} \\ \frac{\text{КЗК}_{ac} + \text{КЗК}_{bd}}{2} & , (ННПС_{ac} = ННПС_{bd}) \wedge (\text{ОСРНП}_{ac} \neq \text{ОСРНП}_{bd}) \\ \text{КЗК}_{ac} + \text{КЗК}_{bd} & , (ННПС_{ac} = ННПС_{bd}) \wedge (\text{ОСРНП}_{ac} = \text{ОСРНП}_{bd}) \end{cases};$$

фвппэ() – функция вычисления параметра похожести элементов контекстного окружения;

$\text{снп}(\text{НП}_{pi}, \text{НП}_{jk})$ – функция определения эквивалентности наименований понятий, причем $\text{снп}(\text{НП}_{pi}, \text{НП}_{jk}) \in \{0,1\}$, НП_{pi} – i-й элемент формализованного смыслового описания рассматриваемого документа, НП_{jk} – k-й элемент формализованного смыслового описания j-ого документа контрольного массива.

Условием глобального смыслового сходства является сходство порядка следования наименований понятий, но, поскольку порядок следования наименований понятий учтен при подсчете коэффициентов $м_{ik}$, с точностью до перестановок слов и словосочетаний, которые возможны в идентичных по смыслу текстах, после этого производится поиск последовательностей наименований понятий, у которых значения локальной смысловой схожести $м_{ik}$ выше некого заданного порога $k_{псх}$. Эта последовательность и будет отрезком текста, для которого выполняется условие глобального смыслового сходства, а мера его выполнения вычисляется как среднее значение характеристик выполнения условия локального смыслового сходства, содержащихся в этих последовательностях наименований понятий. Эта величина и будет являться коэффициентом смыслового сходства фрагментов текстов:

$$k_{сх} = \frac{\sum_{i=0}^{n_{\text{НП}_\text{р}}} \max_k (м_{ik})}{n_{\text{НП}_\text{р}}},$$

где $\max_k (м_{ik})$ - максимальное значение $м_{ik}$, при $k \in [1, n_{\text{НП}_j}]$; $n_{\text{НП}_\text{р}}$ – число элементов в КОДКО рассматриваемого документа; $n_{\text{НП}_j}$ – число элементов в КОДКО j-го документа контрольного массива.

На Рис. 3 проиллюстрирован вышеописанный процесс установления смысловой схожести фрагментов анализируемого документа и документа массива.

3. Реализация программного модуля выявления неявно выраженных заимствований в текстах документов

Разработанный программный модуль выявления неявно выраженных заимствований в текстах документов проектировался таким образом, чтобы была возможность его интеграции практически в любую информационную систему. Архитектура этого комплекса показана на Рис. 4. В соответствии с этой архитектурой программный модуль выявления семантической близости содержания текстов базируется на программно-лингвистической платформе МетаФраз и включает три группы функций:

- Функции выявления неявно выраженных заимствований в текстах, предназначенные для реализации всего цикла автоматической обработки текста, формализации его смыслового представления и обеспечения возможности сопоставления этого представления с аналогичными представлениями других текстов.
- Функции управления и визуализации процесса выявления заимствований в текстах, предназначенные для управления процессами обработки текстов, настройки параметров обработки и визуализации основных этапов и результатов этой обработки. Подсистема включает комплекс графических интерфейсов процессов обработки текстов.
- Функции хранения текстов документов и их формализованных смысловых представлений, предназначенные для обеспечения процессов загрузки, обработки, формализации и хранения текстов и их формализованных смысловых представлений, а также поиска по этим представлениям.

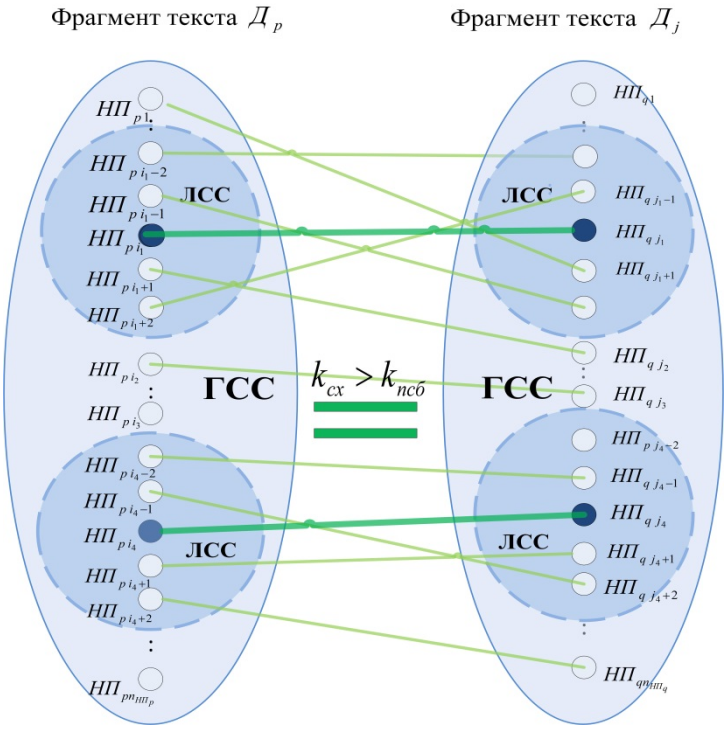

Рис. 3. Установление смысловой схожести фрагментов анализируемого документа и документа массива

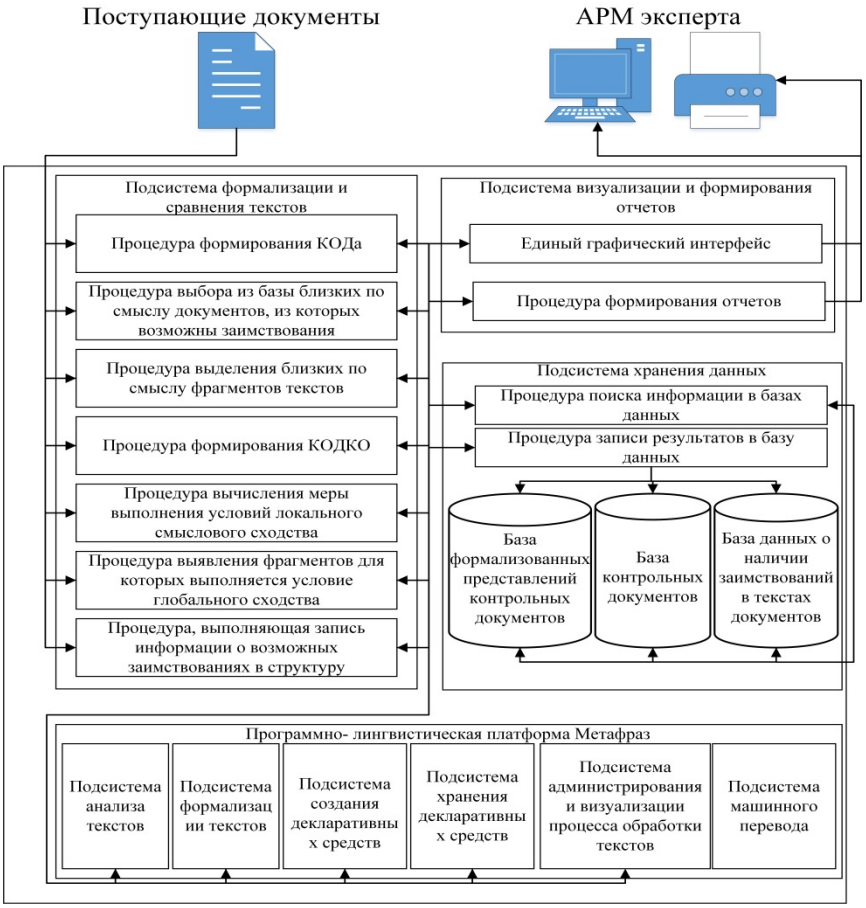

Рис.4. Архитектура модуля выявления семантической близости содержания текстов

Табл.1. Среднее значение полноты и точности выявления заимствований методами ВНВЗ и «шинглов»

Полнота		**Точность**		**F1 – мера**	
метод ВНВЗ	метод «шинглов»	метод ВНВЗ	метод «шинглов»	метод ВНВЗ	метод «шинглов»
0.73	0.45	0.94	0.96	0.82	0.61

Заключение

Для оценки эффективности метода выявления неявно выраженных заимствований и его реализации был проведен эксперимент: разработанный метод был протестирован на коллекции документов размером 5398. В ходе этого эксперимента результаты работы метода выявления неявно выраженных заимствований были сопоставлены с результатами, полученными с применением метода «шинглов», который часто используется при поиске заимствований. Для этого были посчитаны параметры полноты, точности и меры процесса выявления заимствований F1 в текстах документов (Табл. 1).

Анализ результатов эксперимента показал, что значения параметров полноты и F1-меры выявления заимствований в текстах массива научно-технических документов, полученные с помощью разработанного метода в среднем выше на 28 и 21%, чем значения таких параметров, полученные с помощью метода «шинглов» (при длине шингла равной 3). Но при этом наблюдается незначительное снижение (в среднем на 2%) параметра точности выявления заимствований.

Алгоритмы, разработанные в ходе данного исследования, использованы в информационной системе «Мониторинг СМИ», которая функционирует в режиме промышленной эксплуатации. В ее базе данных уже накоплено более 37 млн. документов и ежедневно в нее поступает и оперативно обрабатывается более 100 тыс. документов и новостных сообщений по различным тематикам.

Литература

1. Salton G., Wong A., Yang C.S. 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.
2. 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.
3. Зеленков Ю.Г., Сегалович И.В. Сравнительный анализ методов определения нечетких дубликатов для WEB-документов // Труды 9-й Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» RCDL’2007. Переславль-Залесский. 2007.
4. Broder A. On the resemblance and containment of documents. Compression and Complexity of Sequences (SEQUENCES'97), pages 21-29. IEEE Computer Society, 1998.
5. Broder S., Glassman M., Manasse, Zweig G. Syntactic clustering of the Web. Proc. of the 6th International World Wide Web Conference, April 1997.
6. Палагин А.В., Кривой С.Л., Петренко Н.Г. Концептуальные графы и семантические сети в системах обработки естественно-языковой информации // Математические машины и системы. Киев, 2009. №3.С.67-79
7. Богатырев М.Ю., Латов В.Е., Столбовская И.А. Применение концептуальных графов в системах поддержки электронных библиотек // Труды 9-й Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» RCDL’2007. Переславль-Залесский. 2007. Т. 2. С. 104-110.
8. Hassan S., Mihalcea R. Measuring semantic relatedness using salient encyclopedic concepts// Artificial Intelligence, Special Issue, 2011.
9. 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.
10. Захаров В.Н., Хорошилов А.А. Методы решения задачи автоматического выявления заимствований в структурированных научно-технических документах на основе их семантического анализа // Труды 15-й Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» RCDL’2013. Ярославль. 2013.
11. 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.
12. Лукашевич Н.В. Тезаурусы в задачах информационного поиска. М: Изд. МГУ. 2011.508 с.
13. Захаров В.Н., Хорошилов А.А. Автоматическая оценка подобия тематического содержания текстов на основе сравнения их формализованных смысловых описаний // Труды 14-й Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» RCDL’2012. Переславль-Залесский. 2012.
14. Хорошилов А.А. Методы выявления имплицитно выраженных заимствований в научно-технических текстах на основе их концептуального анализа // Труды 15-й Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» DAMDID/RCDL’2015, Обнинск. 2015 г.

Захаров Виктор Николаевич. Ученый секретарь Федерального государственного учреждения ФИЦ ИУ РАН. Окончил МГУ им. М.В. Ломоносова в 1971 году. Доктор технических наук, доцент. Автор 190 печатных работ из них 10 монографий. Область научных интересов: программное обеспечение для вычислительных и инфокоммуникационных систем и сетей; безопасность, структурная надежность и живучесть вычислительных и инфокоммуникационных систем и сетей; средства создания и поддержки проблемно-ориентированных систем, основанных на знаниях, и экспертных систем; средства создания и поддержки электронных библиотек и электронных изданий. E-mail: vzakharov@ipiran.ru

Хорошилов Александр Алексеевич. Ведущий научный сотрудник Федерального государственного учреждения ФИЦ ИУ РАН. Окончил Кишиневский политехнический институт в 1973 году. Доктор технических наук. Автор 87 печатных работ из них двух монографий. Область научных интересов: компьютерная лингвистика, информационные технологии. E-mail: khoroshilov@mail.ru

Хорошилов Алексей Александрович. Научный сотрудник Федерального государственного учреждения ФИЦ ИУ РАН. Окончил Московский авиационный институт в 2011 году. Кандидат технических наук. Автор 23 печатных работ. Область научных интересов: компьютерная лингвистика, информационные технологии. E-mail: a.a.horoshilov@mail.ru

A method for detecting implicit plagiarism in scientific and technical texts

V.N. Zakharov, A.A. Khoroshilov, A.A. Khoroshilov

Abstract. The paper considers the process of automatic detection of implicit plagiarism in documents on the base of comparison of their formalized representations and calculation of the measures of local semantic similarity of concepts and global semantic similarity of text fragments. In solving this problem we developed a model of the semantic structure of texts and methods for formalization and detection of semantic proximity of the texts under comparison. We also developed the methods for identification of the text fragments similar in semantic structure. The main advantage of this method is that it makes it possible to detect different kinds of plagiarism including the most complex cases of implicit plagiarism. In the study, the results of the work were compared to the results obtained with the use of the method of "shingles". The proposed method showed high efficiency.
Keywords: plagiarism detection, automated text processing, formal description of text, semantic structure, linguistic software, declarative means.

References

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.
2. 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.
3. Zelenkov Yu.G., Segalovich I.V. 2007. Sravnitel'nyy analiz metodov opredeleniya nechetkikh dublikatov dlya WEB-dokumentov [Comparative analysis of near-duplicate detection methods of Web documents]. Rus. Trudy 9-y Vserossiyskoy nauchnoy konferentsii «Elektronnye biblioteki: perspektivnye metody i tekhnologii, elektronnye kollektsii» RCDL'2007 [Digital libraries: advanced methods and technologies, digital collections: works IX of All-Russian National Research Conference Pereslavl], Pereslavl-Zalesskiy, Russia, 2007.
4. Broder On the resemblance and containment of documents. Compression and Complexity of Sequences (SEQUENCES'97), pages 21-29. IEEE Computer Society, 1998.
5. Broder, S. Glassman, M. Manasse and G. Zweig. Syntactic clustering of the Web. Proc. of the 6th International World Wide Web Conference, April 1997.
6. V. Palagin, S. L. Krivoy, N. G. Petrenko. 2009. Kontseptual'nye grafy i semanticheskie seti v sistemakh obrabotki estestvenno-yazykovoy informatsii [Conceptual graphs and semantic network in natural language information processing systems]. Rus. Matematicheskie mashiny i sistemy [Mathematical Machines and Systems]. Kiev, 2009, N 3.-p.67-79
7. M. Bogatyrev, V. Latov, I. Stolbovskaya. 2007. Primenenie kontseptual'nykh grafov v sistemakh podderzhki elektronnykh bibliotek [Application of Conceptual Graphs in Digital Libraries]. Rus. Trudy 9-y Vserossiyskoy nauchnoy konferentsii «Elektronnye biblioteki: perspektivnye metody i tekhnologii, elektronnye kollektsii» RCDL'2007 [Digital libraries: advanced methods and technologies, digital collections: works IX of All-Russian National Research Conference Pereslavl], Pereslavl, Russia, 2007.
8. Hassan S., Mihalcea R. Measuring semantic relatedness using salient encyclopedic concepts// Artificial Intelligence, Special Issue, 2011.

9. 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.
10. Zakharov V.N., Khoroshilov A.A. 2013. Metody resheniya zadachi avtomaticheskogo vyyavleniya zaimstvovaniy v strukturirovannykh nauchno-tekhnicheskikh dokumentakh na osnove ikh semanticheskogo analiza [Semantic methods for solving a problem of automatic detection of plagiarism in structured scientific and technical documents] Rus. Elektronnye biblioteki: perspektivnye metody i tekhnologii, elektronnye kollektsii: Trudy XV Vserossiyskoy nauchnoy konferentsii RCDL'2013 (Yaroslavl', 14–17 oktyabrya 2013) [E-libraries: prospective methods and technology, electronic collections: works XV of All-Russian scientific conference of rcdl'2013 (Yaroslavl ,14-17 October 2013)] Yaroslavl, 30–38.
11. Belonogov G.G. 2008. Teoreticheskie problemy informatiki] Semanticheskie problemy informatiki [Theoretical problems of Informatics. Semantic problems of Informatics]. Rus. Moscow: G. V. Plekhanova REA 2:238.
12. Lukashevich N.V. 2011. Tezaurusy v zadachakh informatsionnogo poiska [thesaurus of information search]. Rus. Moscow: Publishing Moscow State University, 508 p.
13. Zakharov V.N., Khoroshilov A.A. 2012. Avtomaticheskaya otsenka podobiya tematicheskogo soderzhaniya tekstov na osnove sravneniya ikh formalizovannykh smyslovykh opisaniy [Automatic assessment of similarity of the texts' thematic content on the base of their formalized semantic descriptions comparison] Rus. Elektronnye biblioteki: perspektivnye metody i tekhnologii, elektronnye kollektsii: Trudy XIV Vserossiyskoy nauchnoy konferentsii RCDL'2012 [Digital libraries: advanced methods and technologies, digital collections: works XIV of All-Russian scientific conference of rcdl'2012]. Pereslavl,October 15-18, 2012.
14. Khoroshilov A.A. 2015. Metody vyyavleniya implitsitno vyrazhennykh zaimstvovaniy v nauchno-tekhnicheskikh tekstakh na osnove ikh kontseptual'nogo analiza [A Method for Detecting Implicit Plagiarism in Scientific and Technical Texts on the Basis of their Conceptual Analysis] Rus. Trudy XV-oy Vseros. nauch. konf. «Elektronnye biblioteki: perspektivnye metody i tekhnologii, elektronnye kollektsii» [XVII International Conference DAMDID/RCDL'2015 Data Analytics and Management in Data Intensive Domains]. October 13 – 16, 2015, Obninsk, Russia.

Victor Nikolaevich Zakharov. Scientific Secretary, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow, Russian Federation. Doctor of Sciences, Associate Professor. Number of publications: 190 (10 monographs). Research interest: software for computing and communication systems and networks; safety, structural reliability and survivability of computer and communication systems and networks; tools for creating and supporting problem-oriented systems based on knowledge, and expert systems; means of creating and maintaining digital libraries and electronic publications.

Aleksandr Alekseevich Khoroshilov. Leading Researcher, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow, Russian Federation. Doctor of Sciences. Number of publications: 87 (2 monographs). Research interest: computational linguistics, information technologies.

Alexey Aleksandrovich Khoroshilov. Researcher, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow, Russian Federation. Candidat of Sciences. Number of publications: 23. Research interest: computational linguistics, information technologies.