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

Захаров Виктор Николаевич
Хорошилов Алексей Александрович

Актуальность: пример работы существующих систем выявления заимствований

Наименование текста	Фрагмент текста (http://ru.wikipedia.org/wiki/Отечественная_война_1812_года)	Измененный фрагмент текста
Текст	С начала вторжения французской армии на территорию Российской империи в июне 1812 года, русские войска постоянно отступали. Быстрое продвижение и подавляющее численное превосходство французов лишали главнокомандующего русской армии, генерала-от-инфантерии Барклая-де-Толли, возможности подготовить войска к сражению. Затянувшееся отступление вызвало общественное недовольство, поэтому император Александр I сместил Барклая-де-Толли и назначил главнокомандующим генерала-от-инфантерии Кутузова[4]. Однако и новый главнокомандующий избрал путь отступления. Стратегия, выбранная Кутузовым, была основана, с одной стороны, на изнурении противника, с другой — на ожидании подкреплений, достаточных для решающего сражения с армией Наполеона[1].	Как известно, с начала вторжения на территорию Российской империи 24 июня 1812 года французской армии, русские войска все время отступали. ~~Подавляющее~~ численное превосходство и стремительное продвижение французов лишали командующего русскими войсками, ~~генерала-от-инфантерии~~ М. Б. Барклая-де-Толли возможности подготовить к битве войска. Долгое отступление вызвало недовольство общественности, ~~поэтому~~ и после этого Российский император Александр сместил Михаила Богдановича Барклая-де-Толли, а потом назначил главнокомандующим русской армии генерала-от-инфантерии Михаила Илларионовича Кутузова[4], но новый главнокомандующий также избрал путь отступления. Стратегия ~~выбранная~~ Кутузова, основывалась, как, на изнурении противника, так и на ожидании подкреплений, которых было бы достаточно для решающей битвы с армией французского императора[1].
Результат поиска в системе http://www.antiplagiat.ru	**Ссылка на источник http://ru.wikipedia.org/wiki/Бородинское сражение**	**Нет ни одного источника!**

-вставка слов
-изменение порядка слов
-незначительно изменение слов на их смысловые инварианты

Особенности поиска заимствований в научно технических документах (отчеты по НИОКР, диссертации и др.)

- Особая структура документа
- Сравнение документов в пределах рубрики
- Важность получения точных результатов

Принципы построения процедур автоматического выявления заимствований

Программное обеспечение системы автоматического выявления заимствований

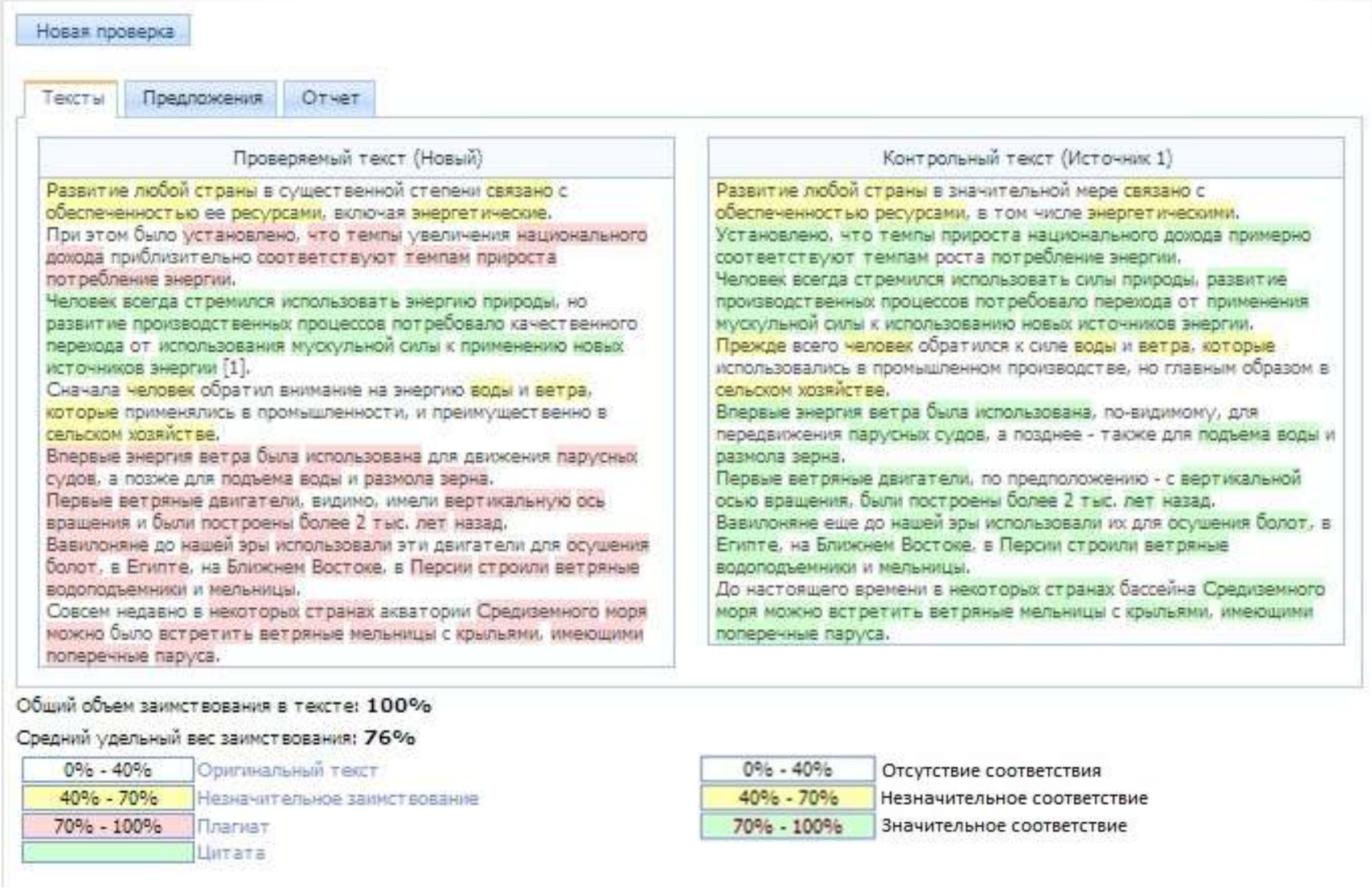
Характеристики для проверяемых текстов:

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

Параметры для каждого совпавшего предложения :

- Номер каждого из сравниваемых предложений в соответствующих текстах.
- Вектор соответствий слов в предложениях
- Текст предложения.

Пример работы программы



Пример работы программы

Новая проверка

Тексты | Предложения | Отчет

Не показывать оригинальные предложения

		Проверяемый текст (Новый)
1	63%	Развитие любой страны в существенной степени связано с обеспеченностью ее ресурсами, включая энергетические.
2	76%	При этом было установлено, что темпы увеличения национального дохода приблизительно соответствуют темпам прироста потребление энергии.
3	89%	Человек всегда стремился использовать энергию природы, но развитие производственных процессов потребовало качественного перехода от использования мускульной силы к применению новых источников энергии [1].
4	53%	Сначала человек обратил внимание на энергию воды и ветра, которые применялись в промышленности, и преимущественно в сельском хозяйстве.
5	80%	Впервые энергия ветра была использована для движения парусных судов, а позже для подъема воды и размола зерна.
6	86%	Первые ветряные двигатели, видимо, имели вертикальную ось вращения и были построены более 2 тыс. лет назад.
7	87%	Вавилоняне до нашей эры использовали эти двигатели для осушения болот, в Египте, на Ближнем Востоке, в Персии строили ветряные водоподъемники и мельницы.
8	75%	Совсем недавно в некоторых странах акватории Средиземного моря можно было встретить ветряные мельницы с крыльями, имеющими поперечные паруса.

	Контрольный текст (Источник 1)	
1	Развитие любой страны в значительной мере связано с обеспеченностью ресурсами, в том числе энергетическими.	
2	Установлено, что темпы прироста национального дохода примерно соответствуют темпам роста потребление энергии.	
3	Человек всегда стремился использовать силы природы, развитие производственных процессов потребовало перехода от применения мускульной силы к использованию новых источников энергии.	89%
4	Прежде всего человек обратился к силе воды и ветра, которые использовались в промышленном производстве, но главным образом в сельском хозяйстве.	
5	Впервые энергия ветра была использована, по-видимому, для передвижения парусных судов, а позднее - также для подъема воды и размола зерна.	
6	Первые ветряные двигатели, по предположению - с вертикальной осью вращения, были построены более 2 тыс. лет назад.	
7	Вавилоняне еще до нашей эры использовали их для осушения болот, в Египте, на Ближнем Востоке, в Персии строили ветряные водоподъемники и мельницы.	
8	До настоящего времени в некоторых странах бассейна Средиземного моря можно встретить ветряные мельницы с крыльями, имеющими поперечные паруса.	

Общий объем заимствования в тексте: **100%**

Средний удельный вес заимствования: **76%**

0% - 40%	Оригинальный текст
40% - 70%	Незначительное заимствование
70% - 100%	Плагиат
	Цитата

Модель представления текста

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

Отождествление наименований понятий

- Обобщение на уровне словоизменения
- Обобщение на уровне словообразования
- Обобщение на уровне синонимии

Проверка разработанного программного обеспечения

Проверка на заимствования документа с именем 0220xxxxx84 (921 предложение)

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

Применение систем установления заимствований

- Школы
- ВУЗы
- Электронные библиотеки
- Диссертационные советы

Развитие систем установления заимствований

- Повышение точности обработки за счет создания тематических концептуальных словарей большего количества различных предметных областей
- Увеличение производительности за счет использования за счет использования технологий параллельной обработки
- Создание удобных инструментов для эксперта-аналитика, использующего средства установления заимствований

Спасибо за внимание