# Автоматическая оценка подобия тематического содержания текстов на основе сравнения их формализованных смысловых описаний

© В.Н. Захаров
ИПИ РАН

© А. А. Хорошилов
ЦИТиС

Москва

vzakharov@ipiran.ru

a.a.horoshilov@mail.ru

**Аннотация**

В работе описываются методы автоматического построения формализованного смыслового описания документа и оценки подобия тематического содержания текстов. Эти методы базируются на применении процедур семантико-синтаксического и концептуального анализа, обеспечивающих выявление понятийного состава текста и назначения наименованиям понятий характеристик, соответствующих их семантической роли и значимости в тексте. Автоматическая оценка подобия тематического содержания текстов производится путем сравнения понятийного состава текстов. Результатом работы явилось создание комплекса программных средств, предназначенного для оценки подобия тематического содержания текстов. Основными преимуществами данного программного комплекса является его быстродействие и возможность обработки текстов, относящихся к любым предметным областям.

## 1 Введение

### 1.1 Проблемы обработки текстовой информации

В настоящее время в связи с постоянно растущими объемами информационных ресурсов доступ пользователей к интересующим их сведениям становится все более затруднительным. Для решения этой проблемы создаются современные информационные технологии, базирующиеся на мощном фундаменте телекоммуникационных и вычислительных средств. Сейчас эти средства достигли высокого уровня развития. Особенно ярко эти успехи проявляются в области развития средств связи и разработки мощных вычислительных систем.

На фоне этих успехов успехи в области смысловой обработки информации менее значительны. Эти успехи зависят, прежде всего, от достижений в изучении процессов человеческого мышления, процессов речевого общения между людьми и от умения моделировать эти процессы на ЭВМ. Основной проблемой, возникающей при обработке текстовой информации, является трудность автоматического составления формализованного описания смыслового содержания документов и, как следствие этого - трудность установления смысловой связи между различными документами. Это обусловлено тем, что в разных текстах одни и те же ситуации могут описываться в терминах различной степени общности и с помощью различных языковых средств. И только человек, анализирующий документы, руководствуясь своими представлениями о содержании документов и средствах выражения этого содержания и опираясь на свои профессиональные знания и опыт, в состоянии установить степень смысловой близости анализируемых документов. Большинство систем автоматической обработки текстовой информации, функционирующих в настоящее время, не могут в полной степени решать эти проблемы.

В связи с этим возникает необходимость в разработке эффективных методов автоматического анализа содержания документов. Отличительная особенность предлагаемых методов заключается в том, что они базируются на современных представлениях о смысловой структуре текстов и оригинальных процедурах семантико-синтаксического и концептуального анализа.

### 1.2 Методы сравнения текстов

В рамках наших исследований были рассмотрены такие методы, как TF, Opt Freq, Lex Rand, Log_Shingle, Megashingles, Long Sent, Descr Words [11-15]. Широкомасштабный эксперимент по сравнительному анализу этих методов был выполнен Ю.Г. Зеленковым и И.В. Сегаловичем в работе [11]. В этой работе ставилась задача оценить качество наиболее известных, разнообразных и эффективных с вычислительной точки зрения алгоритмов определения нечетких дубликатов. При этом предполагалось сравнивать алгоритмы по параметрам полноты и точности, а также определить их взаимную корреляцию и совместное покрытие разными сочетаниями алгоритмов исходного мно-

**Труды 14-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2012, Переславль-Залесский, Россия, 15-18 октября 2012 г.**

жества пар нечетких дубликатов. В качестве тестового массива использовалась веб-коллекция документов РОМИП (около 500 тыс. документов).

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

Необходимо отметить, что во всех рассмотренных алгоритмах текст рассматривается как некоторое множество, состоящее из отдельных слов. Различные операции, выполняемые в процессе поиска текстов-дубликатов, производились над словам и их цепочкам. Между тем текст это не множество слов и их последовательностей, и при установлении смысловой близости документов нужно сопоставлять, прежде всего, смысловые единицы текста – понятия, выраженные словосочетаниями. При этом, необходимо учитывать такое явление как вариативность форм представления в тексте одного и того же смысла. (См. ниже). А это явление в вышерассмотренных алгоритмах полностью игнорировалось. Поэтому алгоритмы установления смысловой близости документов должны базироваться на современных процедурах семантико-синтаксического и концептуального анализа, позволяющих выявлять в текстах наименования понятий, представленные в различных формах их представлений.

## 2 Процедуры семантического анализа документов

### 2.1 Единицы языка и речи

Основными единицами языка и речи, принятыми в лингвистике, являются морфемы, слова, словосочетания, фразы и различного рода сверхфразовые единства. Система единиц языка и речи обычно представляется в виде иерархической структуры, в которой единицы вышестоящих уровней включают в свой состав единицы нижестоящих уровней и сами входят в состав единиц более высоких уровней. Для каждого уровня единиц языка разработаны инструментальные средства их обработки. Для обработки слова обычно используется морфологический анализ. Для обработки предложений и сверхфразовых единств (текстов) обычно применяется семантико-синтаксический и концептуальный анализ.

### 2.2 Морфологический анализ

Морфологический анализ слов естественных языков предназначен для определения структуры слов и назначения им грамматических признаков. Используемый в наших исследованиях морфологический анализ разработан профессором Г. Г. Белоноговым [5] на основе созданной им системы флективных классов русских слов. Система флективных классов была создана путем анализа текстов, в которых в различных контекстных окружениях слова могут приобретать различные формы. Это могут быть формы словоизменения и словообразования.

Процедура морфологического анализа функционирует следующим образом. На первом этапе производится поиск в словаре "служебных и коротких слов", а затем, в случае неудачи, в словаре концов словоформ. Результаты анализа, полученные в процессе поиска по первому словарю, считаются правильными. Вероятность правильного анализа слов по словарю концов словоформ при обработке текстов любой тематики превышает 99% [4,5].

### 2.3 Семантико-синтаксический анализ

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

Алгоритм синтаксического анализа текстов, как и множество подобных ему алгоритмов, имеет тот недостаток, что в нем в явном виде не выделяются смысловые единицы, выраженные словосочетаниями. В свою очередь смысловое содержание текстов документов выражается с помощью единиц смысла – понятий и связей между ними. Профессор Г. Г. Белоногов [4,5] определяет 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. Под устойчивыми фразеологическими словосочетаниями мы будем понимать не только идиоматические выражения и терминологические словосочетания, но и любые повторяющиеся отрезки связных текстов, для их выделения применяется процедура концептуального анализа.

### 2.4 Концептуальный анализ

Процедура концептуального анализа текстов предназначена, прежде всего, для выявления наименований понятий в тексте. Эта процедура базируется на результатах семантико-синтаксического анализа и использовании эталонного словаря наименований понятий предметной области, к которой принадлежит анализируемый текст.

На первом этапе текст обрабатывается программами семантико-синтаксического анализа, в результате которого текст членится на предложения, строится синтаксическая структура предложений, и каждому слову назначается набор грамматических признаков. Затем каждое предложение исходного текста разбивается на различные фрагменты и на их основе формируются "поисковые образы" в виде последовательностей нормализованных слов и словосочетаний. Далее эти последовательности заменяются на их первичные хеш-коды – на более короткие восьмибайтовые кодовые комбинации, которые в дальнейшем используются в процессе отождествления отрезков исходного текста с наименованиями понятий эталонного словаря.

После того как текст был представлен в виде списка слов и словосочетаний, из него выбираются наиболее информативные слова и словосочетания. Такой выбор осуществляется по эталонному словарю наименований понятий (концептуальный анализ с контролем по тезаурусу) или путем проверки структуры словосочетаний программой синтаксического контроля и последующего исключения из их состава малоинформативных словосочетаний по так называемому словарю стоп-слов.

## 3 Концепция смысловой обработки текстовой информации

### 3.1 Структура языка и речи

При разработке процедур автоматической обработки текстовой информации важно исходить из правильных представлений о смысловой структуре языка и речи. По современным представлениям наиболее информативными и наиболее устойчивыми единицами смысла являются понятия [4-6,9]. 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 более высоких уровней. Второй по значимости единицей смысла является предложение. Из предложений формируются различного рода сверхфразовые единства, которые представляются в виде последовательностей предложений связного текста.

Основной чертой предложений является их предикативность – то есть то их свойство, что в них утверждается наличие у объектов определенных признаков и их отношений [4,5]. Свойством предикативности обладают и высказывания, формулируемые на формализованных языках. Таким образом, в основе и предложений на естественном языке, и формализованных логических высказываний лежит предикатно-актантная структура, компонентами которой являются понятия-предикаты (отношения) и понятия-актанты, выступающие в роли описываемых объектов.

В соответствии с положенной в основу наших исследований концепцией в текстах понятия-актанты выражаются чаще не отдельными словами, а устойчивыми словосочетаниями. А словами или словосочетаниями, устанавливающими смысловые отношения между ними - понятиями-предикатами - являются обычно глаголы или отглагольные формы существительных, прилагательных и наречий. При этом необходимо учитывать, что в текстах 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 наименований понятий.

Таким образом, при проведении исследований необходимо выявление понятийной структуры текста. Под такой структурой текста будем понимать совокупность понятий, выявленных в тексте и связанных между собой смысловыми отношениями. Между тем выявленную понятийную структуру текста, состоящую из текстовых форм наименований понятий, необходимо автоматически привести к формализованной форме ее представления. Такое приведение выполняется путем автоматической нормализации текстовых форм наименований понятий (слов или словосочетаний) к их каноническим формам.

### 3.2 Приведения понятий к нормализованной (канонической) форме

Обычно под нормализованной (канонической) формой слова понимается та его форма, которая традиционно указывается в словарях. Например, для существительного - это форма именительного падежа единственного или (в случае pluralia tantum) множественного числа, для глагола – форма инфинитива, для прилагательного – форма именительного падежа единственного числа мужского рода. Процедура замены исходной вариантной формы слова на каноническую называется процедурой нормализации или лемматизации.

Необходимо отметить, что нормализация слов/словосочетаний может выполняться с различной степенью смысловой общности – на уровне словоизменения или на уровне словообразования. Порядок слов в словосочетании и неизменяемые формы слов при нормализации не изменялись.

### 3.3 Концепция формализованного смыслового описания документа

Исходя из вышесказанного, смысловую структуру текста можно представить в виде совокупности нормализованных наименований понятий и связей между ними. Такую смысловую структуру текста будем называть его формализованным смысловым описанием.

В состав формализованного смыслового описания документа должны быть включены наименования понятий, сопровождаемые коэффициентом, определяющим степень их смысловой значимости в тексте. Поэтому при формировании формализованного описания документа нужно определить его состав и назначить каждому элементу его весовой коэффициент. Для этого необходимо в анализируемом тексте выявить информативные слова или словосочетания, опираясь на их формальных характеристиках, среди которых: значения их частот в предметной области и конкретном тексте, длины словосочетаний (в словах), принадлежности к категории географических названий или фамильно-именной группе, а также их наличие или отсутствие в эталонном концептуальном словаре и их наличие или отсутствие в словаре стоп-слов.

В формализованном смысловом описании документа каждый элемент состоит из пары наименований понятий-актантов, связанных между собой понятием-предикатом.

Таким образом, можно сформулировать следующее 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.

### 3.4 Определение весовых коэффициентов наименований понятий

Для указания смысловой значимости наименования понятия в формализованном смысловом описании документа необходимо назначить каждому наименованию понятия весовой коэффициент.

На весовой коэффициент оказывают влияние следующие характеристики наименования понятия:

- Значение числа встречаемости наименования понятия в предметной области (глобальная частота)
- Значение числа встречаемости наименования понятия в тексте (локальная частота)
- Длина наименования понятия (в словах)
- Принадлежность наименования понятия к фамильно-именной группе.

При назначении весовых коэффициентов мы воспользуемся предложенной нами формулой:

$$W_{ij} = \begin{cases} (p_{ij} + fg_{ij}) \cdot f_{ij} \cdot l_{ij} & l_{ij} \le k_{\max} \\ (p_{ij} + fg_{ij}) \cdot f_{ij} \cdot k_{\max} & l_{ij} > k_{\max} \end{cases},$$

где, $p_{ij}$ - коэффициент, увеличивающий степень значимости наименования понятия в зависимости от его принадлежности к фамильно-именной группе, географическим названиям и т.д.

$l_{ij}$ - количество слов в словосочетании, которым выражается j-ое понятие в i-ом тексте;

$f_{ij}$ - частота появления j-ого понятия в i-ом тексте;

$fg_{ij}$ - нормированная глобальная частота j-ого понятия в i-ом тексте;

$k_{\max}$ - коэффициент, установленный опытным путем, соответствующий максимальной длине словосочетания, после которой она не должна влиять на итоговый вес наименования понятия.

### 3.5 Порядок построения табличного представления формализованного смыслового описания документа

На основании высказанного можно определить следующий порядок построения табличного представления формализованного смыслового описания документа:

1. Определение синтаксической и концептуальной структуры текста;
2. Разрешение анафорических ссылок в тексте;
3. Получение частотного словаря наименований понятий;
4. Установление смысловых связей между наименованиями понятий;
5. Исключение малоинформативных слов или словосочетаний;
6. Приведение различных форм представления наименований понятий к единой унифицированной форме;
7. Дополнение полученной по тексту таблицы связей наименований понятий внеконтекстными парадигматическими и ассоциативными связями.

## 4 Автоматическая оценка подобия тематического содержания текстов

### 4.1 Описание процесса автоматической оценки подобия тематического содержания текстов

Процедура автоматической оценки подобия тематического содержания текстов выполняется путем сопоставления формализованного смыслового содержания двух документов. Поскольку в данной работе задача состоит в определении смысловой близости тематически связанных документов, в которых освещены те же темы, но возможно в несколько другом аспекте, мы можем упростить формализованное смысловое описание и в данном случае исключить из него связи между объектами. Тогда формализованное смысловое описание документа примет следующий вид:

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

$n_F$ - количество элементов в формализованном смысловом описании документа;

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

$Nc_i$ — наименование понятия;

$w_i$ - вес наименования понятия.

Для выполнения данной задачи оценки подобия тематического содержания текстов необходимо установить формальные критерии, определяющие численную характеристику степени их подобия. Эта характеристика получена как произведение коэффициента нормировки и частного от деления суммы весов совпавших наименований понятий на сумму весов всех наименований понятий эталонного документа. Назовем эту характеристику коэффициентом подобия тематического содержания текстов. Тогда формулу для вычисления коэффициента подобия тематического содержания p-ого и q–ого текстов можно записать следующим образом:

$$K_{\text{sim}} = \frac{\sum_{j=1}^{n_\cap} w_{\cap j} \cdot \sum_{j=1}^{n_p} f_{pj}}{\sum_{j=1}^{n_p} w_{pj} \cdot \sum_{j=1}^{n_q} f_{qj}}$$

$w_{\cap j}$ - j-ая компонента вектора весовых коэффициентов наименований понятий, содержащихся в обоих текстах, причем веса берутся из формализованного смыслового описания q-ого текста.

$w_{pj}$ - j-ая компонента вектора весовых коэффициентов наименований понятий, содержащихся в p-ом тексте.

$f_{pj}$ - j-ая компонента вектора локальных частот наименований понятий, содержащихся в p-ом тексте.

$f_{qj}$ - j-ая компонента вектора локальных частот наименований понятий, содержащихся в q-ом тексте.

$n_\cap$ - размерность вектора наименований понятий, содержащихся в обоих текстах.

$n_p$ - размерность вектора наименований понятий, содержащихся в p-ом тексте.

$n_q$ - размерность вектора наименований понятий, содержащихся в q-ом тексте.

С ростом коэффициента $K_{\text{sim}}$ увеличивается степень тематического подобия тематического содержания текстов. Если коэффициент $K_{\text{sim}}$=1, тогда тексты идентичны.

### 4.2 Пример работы данного алгоритма на коротких запросах

Приведем пример одного из экспериментов, иллюстрирующих работу вышеизложенных алгоритмов, и заключающихся в сравнении результатов ранжирования документов, выполненных поисковой системой Google, программным комплексом, разработанным авторами и человеком-экспертом. Для этого с помощью поисковой системой Google был выполнен поиск по следующему короткому запросу: “Як-38 - самолет ОКБ Яковлева с технологией вертикального взлета”. Далее первые 100 из найденных поисковой системой документов были обработаны с помощью разработанного авторами программного комплекса и отсортированы по степени их подобия поисковому запросу. И, наконец, эти найденные документы также были оценены человеком по степени их релевантности данному поисковому запросу. Результаты этого эксперимента приведены в таблице 1.

В первом столбце отображено ранжирование документов, полученное поисковой системой Google. Во втором столбце – ранжирование, полученное после обработки программным комплексом авторов, и в третьем столбце – ранжирование, полученное экспертом. В четвертом столбце приведена оценка тематического подобия документов запросу, полученная с помощью созданного программного обеспечения, результаты приведены в процентах. Последняя колонка содержит краткие комментарии к анализируемым текстам. Результаты этого эксперимента показывают, что ранжирование документов, выполненное разработанным программным комплексом, более точно соответствует ранжированию, выполненному экспертом-человеком, чем произведенное системой Google.

Таблица 1
*Фрагмент результатов сравнения ранжирований документов по степени их релевантности*

| Ранжирование документов по степени их релевантности поисковому запросу | | |
|---|---|---|
| Поисковая система Google | Человек-эксперт | Программный комплекс авторов |
| 1 | 6 | 8 |
| 2 | 8 | 5 |
| 3 | 9 | 9 |
| 4 | 1 | 1 |
| 5 | 4 | 3 |
| 6 | 11 | 11 |
| 7 | 7 | 7 |
| 8 | 5 | 6 |
| 9 | 10 | 10 |
| 10 | 3 | 4 |
| 11 | 2 | 2 |

## 5 Заключение

Предложенные в данной работе методы были реализованы в виде программного комплекса, и их эффективность была проверена на 25 выборках по 30 текстов разной степени тематического подобия общественно-политической тематики с участием экспертов, оценивающих степень подобия текстов, также данный эксперимент был повторен при аналогичных условиях для текстов по ядерной физике. Проведенные эксперименты подтвердили эффективность изложенных методов. Предлагаемые методы могут использоваться в системах автоматической обработки текстовой информации. В настоящее время данный программный комплекс функционирует в составе Системы оперативного мониторинга СМИ (СКЦ РосАТОМ).

Следующим этапом в развитии этого программного комплекса может быть реализация новых функциональных возможностей:

1. Выделение в тексте смысловых фрагментов, отражающих различные темы документа;
2. Установление смысловой близости документов с учетом связей между объектами;
3. Установление заимствований из других документов;
4. Оценка тождественности (аутентичности) смыслового содержания разноязычных текстов.

## 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] Золотова Г.А. «Коммуникативные аспекты русского синтаксиса» — М., КомКнига, 2010. – 368 с

[4] Белоногов Г.Г., и др. Компьютерная лингвистика и перспективные информационные технологии. Теория и практика построения систем автоматической обработки текстовой информации — М.: Русский мир, 2004. – 264 с.

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

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

[7] Киселев М.В. Метод кластеризации текстов, основанный на попарной близости термов, характеризующих тексты, и его сравнение с метрическими методами кластеризации. – ИНТЕРНЕТ-МАТЕМАТИКА 2007, Екатеринбург: Изд-во Урал. Ун-та, 2007. – 224с.

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

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

[10] 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.

[11] 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.

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

[13] 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.

[14] Д. Гасфилд. Строки, деревья и последовательности в алгоритмах. СПб.: Невский диалект, 2003.

[15] 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.

## Automatic assessment of similarity of the texts' thematic content on the base of their formalized semantic descriptions comparison

Victor Zakharov, Alexey Khoroshilov

The paper describes the methods for automatic generation of the formalized semantic document description and the assessment of the thematic text content similarity. These methods are based on the use of semantic-syntactic and conceptual analysis procedures providing the identification of the conceptual text content and the assignment of the characteristics to the concept names, corresponding to their semantic role in the text. Automatic thematic text content similarity assessment is made by comparison of the conceptual text content.