УДК 004.021:004.912

# ПРОЕКТИРОВАНИЕ И РАЗРАБОТКА АЛГОРИТМА ДЛЯ ПОВЫШЕНИЯ ЭФФЕКТИВНОСТИ СЕМАНТИЧЕСКОГО ПРЕОБРАЗОВАНИЯ ПОИСКОВЫХ ЗАПРОСОВ

**Мельникова Д.А., Рыбанов А.А., Короткова Н.Н.**

*Волжский политехнический институт, филиал ФГБОУ ВПО «Волгоградский государственный технический университет», Волжский, e-mail: dashuta.004@mail.ru*

Проведено исследование существующих алгоритмов семантического преобразования поисковых запросов, основных компонентов шаблонов поиска и алгоритмов машинного обучения. Разработан и предложен алгоритм структурного анализа Web-страницы, предназначенный для увеличения эффективности применения алгоритмов поиска значимой информации. На основе анализа можно реализовать два варианта поиска: классический и поиск похожих текстов. Актуальность заключается в необходимости повышения эффективности методов извлечения знаний. Проводя обзор существующих методов и программных решений, выяснилось, что лидеров в области применения семантического анализа на практике является компания Pearson Knowledge Technologies. Математическое описание представляет собой слова и абзацы с помощью алгоритма и моделирования восприятия текста человеком. В заключении описаны алгоритмы и программные решения первичной обработки.

**Ключевые слова: семантическое преобразование, поиск, алгоритм интеллектуального поиска**

# DESIGN AND DEVELOPMENT OF THE ALGORITHM FOR IMPROVING THE EFFICIENCY OF THE SEMANTIC TRANSFORMATION OF SEARCH QUESTIONS

**Melnikova D.A., Rybanov A.A., Korotkova N.N.**

*Volzhskiy Polytechnical Institute, branch of the Volgograd State Technical University, Volzhskiy, e-mail: dashuta.004@mail.ru*

The study of existing algorithms for semantic transformation of search queries, the basic components of the search templates and machine learning algorithms. Developed and proposed an algorithm of structural analysis of Web page, designed to increase the effectiveness of the algorithms for finding relevant information. Based on the analysis, you can implement two variants of search: the classic search for similar texts. The urgency lies in the need to increase the effectiveness of the methods of knowledge extraction. Conducting a review of existing methods and software solutions, it became clear that leaders in the field of semantic analysis in practice is Pearson Knowledge Technologies. The mathematical description represents the words and paragraphs with an algorithm and modeling of comprehension for the man. In conclusion the described algorithms and software solutions for primary processing.

**Keywords: semantic transformation, search, algorithm of intellectual search**

Большинство современных систем отбора информации из текстов построены для решения определенных задач, сформулированных заранее. Нет системы, которая была бы рассчитана на максимальное решение задачи извлечения – полный синтаксический анализ, семантический анализ произвольного текста и получение всей информации, содержащейся в этом тексте. Любая задача для системы получения данных из текста может быть сформулирована в диапазоне от минимальной до максимальной.

## Постановка проблемы

Целью данной статьи является исследование алгоритмов семантического преобразования поисковых запросов и разработка алгоритма для повышения их эффективности.

Объектом данного исследования являются алгоритмы интеллектуального анализа данных.

Предметом исследования является эффективность алгоритмов интеллектуального анализа содержания Web-ресурсов.

Актуальность статьи заключается в необходимости повышения эффективности методов извлечения знаний на основе информации, которая может извлекаться из распределенных разносторонних источников, что позволит интеллектуализировать Web-контент, дополнив содержание Интернет-ресурсов такими формальными описаниями, которые предоставят возможность интеллектуальным программам или агентам автоматически убеждаться в достоверности определенного контента.

## Обзор существующих методов и программных решений семантичкого преобразования поисковых запросов

Одним из методов семантического преобразования поисковых запросов является сравнение всех контекстов, в которых слова или группы слов употребляются, и контекстов, в которых они не используются, что позволяет сделать вывод о степени близости содержания этих слов или групп слов.

Впервые метод семантического анализа был описан в работе «An Introduction to Latent Semantic Analysis «Landauer, T.K., Foltz, P., and Laham, D. в 1998 году и затем развит в работах Scott Deerwester, Susan Duniais, George Furnas [3].

На данный момент лидером в области применения семантического анализа на практике является компания Pearson Knowledge Technologies. Их коммерческие продукты позволяют убедиться в эффективности метода семантического. Однако, конкретные алгоритмы реализации этого метода не опубликованы, поскольку это является коммерческой тайной, поэтому очень важной частью работы является создание программного обеспечения, которое позволит воспроизвести и проверить результаты работы алгоритмов семантического анализа.

Рассмотрим наиболее распространённые программные продукты для осуществления преобразования поисковых запросов [1]:

*ConceptNet* – семантическая нейронная сеть для понимания текста, написанного человеком;

*FrameNet* – это лексикографический онлайн ресурс, фундаментом которого является идея «фреймовой сементики».

*WordNet* – онлайн семантическая сеть с использованием различных связей слов: лексической, антонимической, контекстной и другими.

### Математическое описание алгоритма поиска значимой информации

Представление слова и абзаца с помощью алгоритма семантического моделирует восприятие текста человеком (рис. 1). Например, с его помощью можно оценить эссе на соответствие теме или сравнить содержание отрывков текста.

Результаты метода достаточно достоверно отражают смысловые корреляции между словами и отрывками.

В качества начальных данных в общей основе алгоритмов семантического анализа используется частота нахождения каждого слова отдельно в тексте, а не частота нахождения каждой фразы.

К примеру, запрос «Apple computer» будет интерпретирован как принадлежность к компьютерной компании Apple, поэтому среди результатов будут документы, которые описывают технологии этой компании (даже если эти документы не содержат слов «Apple» или «computer»).

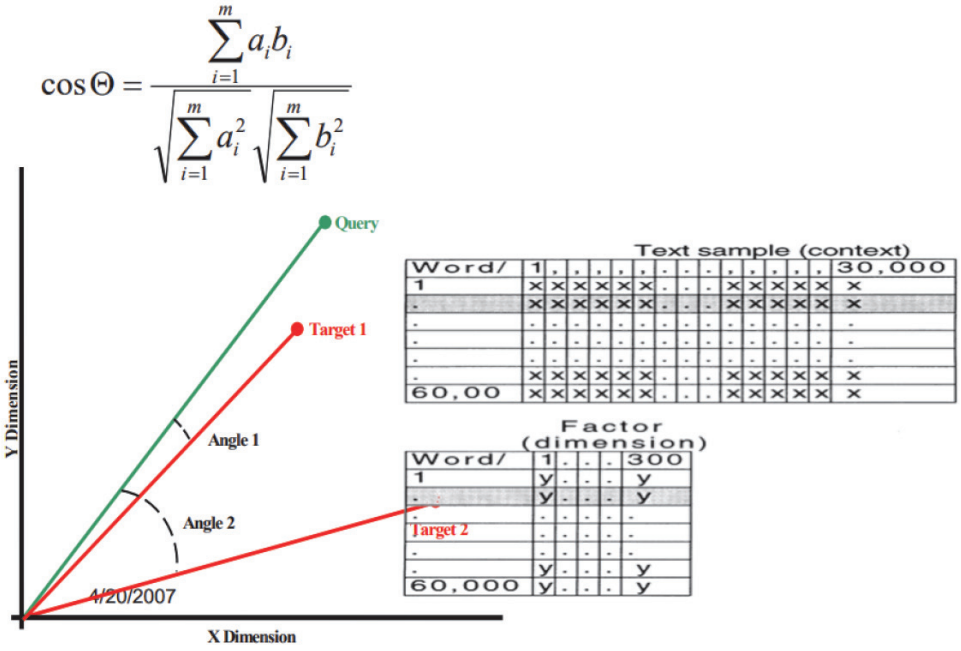


*Рис. 1. Алгоритм семантического преобразования поисковых запросов*

Пусть столбцы матрицы будут означать тексты, а строки – слова. Только элементы матрицы представляют собой количество нахождений конкретного слова в данном тексте. Именно такой подход и является стандартным для семантических моделей.

Итак, матрица отражает связи между словами и текстами. При анализе текстов возникает две проблемы:

• синонимия – когда разные слова обозначают одно и то же понятие;

• полисемия – когда одно и то же слово или несколько слов могут означать различные понятия.

Разработка алгоритма для повышения эффективности семантического преобразования поисковых запросов

На основе алгоритмов семантического анализа можно реализовать два варианта поиска. Первый, более классический, то есть пользователь вводит поисковый запрос, обычно это определенные ключевые слова или фразы, и как результат поиска получает документы, отсортированные в порядке релевантности запросу. Второй вариант реализует поиск похожих текстов. На вход пользователь подает документ, а в результате получает документы, которые похожи на исходный. Поиск, по ключевым словам, происходит среди файлов, которые находятся в определенном каталоге. По этим файлам строится матрица применяемости, но перед этим слова обрабатываются, чтобы уменьшить размерность матрицы.

Понятие «полноты» при отборе информации показывает отношение числа релевантных документов к числу всех найденных документов. В свою очередь «точность» определяется отношением числа найденных релевантных документов к общему числу релевантных документов. Для одновременного учета полноты и точности существует их комбинация (1), которая называется F-показатель и имеет следующий вид [2]:

$$F = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}, \qquad (1)$$

где $P$ – точность; $R$ – полнота; $\beta$ – коэффициент баланса между оценками.

Если $\beta = 1$, то оба показателя имеют равный вес.

Все слова проверяются на принадлежность к так называемому списку «стоп-слов», наиболее часто встречающихся во всех возможных текстах (не только среди тех, по которым будет осуществляться поиск). Они никоим образом не будут отличать документ, текущий от других релевантных.

В результате, после обработки матрицы применяемости, на конвейер анализатора попадают словоформы без суффиксов и окончаний, что дает возможность не учитывать род и падеж слов (рис. 2).

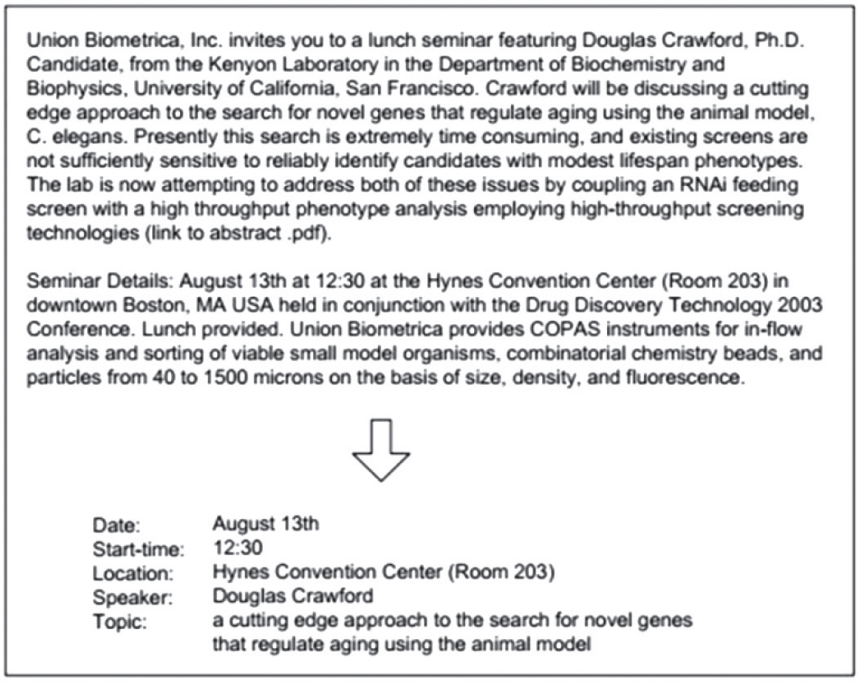


*Рис. 2. Отбор информации из объявления о семинаре*

## Заключение

Описаны алгоритмы и существующие программные решения первичной обработки данных в системах отбора информации, применяемые для определения признаков слов или ключевых запросов в заданных документах.

С целью повышения эффективности работы алгоритмов поиска разработан метод автоматического разделения HTML-страниц на навигационную и содержательную части. Данный алгоритм рекомендуется к применению для узких задач информационного поиска, когда известно, что элементы оформления группы индексированных ресурсов мешают информационному поиску.

## Список литературы

1. Popova S.V. Ranking in keyphrase extraction problem: is it suitable to use statistics of words occurrences? / S.V. Popova, I.A. Khodyrev // Trudy ISP RAN [The Proceedings of ISP RAS]. – 2014. – vol. 26, issue 4. – P. 123–135. (Scopus)

2. Басипов, А.А. Семантический поиск: проблемы и технологии / Басипов А.А., Демич О.В. // Вестник Астраханского государственного технического университета. Серия: Управление, вычислительная техника и информатика. – 2012. – № 1. – С. 104–111.

3. Кравченко, Ю.А. Семантический поиск в SEMANTIC WEB / Кравченко Ю.А., Марков В.В., Новиков А.А. // Известия Южного федерального университета. Технические науки. – 2016. – Раздел II. – С. 65–75.