DOI: 10.17117/na.2016.07.01.407

<http://ucom.ru/doc/na.2016.07.01.407.pdf>

Поступила (Received): 19.07.2016

Куликов Г.С., Курышов Д.М., Солнцев С.П. Семантическое хранилище информационных ресурсов

Kulikov G.S., Kuryshov D.M., Solntsev S.P.
Semantic storage for information resources

В статье рассматривается пример реализации хранилища информационных ресурсов, основанного на семантических принципах. Подробно рассмотрен один из алгоритмов его формирования. Кроме того, в материале представлена основная структура хранилища, положительные и отрицательные стороны его использования

Ключевые слова: семантический, хранилище, информационный, ресурс

The article provides semantic-based realization of information resources storage example. One of the generation algorithms is considered in the article. Moreover, the article contains general structure of the storage, it's advantages and disadvantages

Key words: semantic, storage, information, resource

Куликов Георгий Сергеевич

Бакалавр

Московский технологический университет

г. Москва, пр. Вернадского, 78

Kulikov Georgiy Sergeevich

Bachelor

Moscow technological university

Moscow, Vernadskogo ave., 78

Курышов Денис Михайлович

Бакалавр

Московский технологический университет

г. Москва, пр. Вернадского, 78

Kuryshov Denis Mikhailovich

Bachelor

Moscow technological university

Moscow, Vernadskogo ave., 78

Солнцев Сергей Павлович

Бакалавр

Московский технологический университет

г. Москва, пр. Вернадского, 78

Solntsev Sergey Pavlovich

Bachelor

Moscow technological university

Moscow, Vernadskogo ave., 78

Реляционный подход организации хранения документов имеет ряд недостатков, таких как отсутствие гибкости и динамически изменяющейся структуры, а также невозможность определения связности документов без привлечения дополнительных инструментов и средств.

В данном материале предлагается вариант построения хранилища информационных ресурсов на основе семантики хранимых документов. Использование семантического подхода позволит расширить полигон для анализа данных документа, а также повысит качество описания ресурсов.

Семантика документа

Семантика документа – основной смысл, который передаёт его текст. В данной статье в приоритет ставится автоматическое определение семантики ресурса.

В настоящее время существует достаточно большое количество различных алгоритмов определения смысла текста. Но чаще всего, все они сводятся к тому или иному анализу частотного словаря слов [1].

Алгоритм определения смысла текста

Алгоритм определения смысла текста, используемый при подготовке данной статьи, состоит из следующих шагов:

- подготовка документа;
- формирование частотного словаря слов;
- анализ полученных данных.

Подготовка документа

Прежде чем приступить к анализу документа текст необходимо подготовить. Сначала документ проверяется на возможность анализа. Иными словами, нужно ответить на вопрос, можно ли выделить из него текст (Такие форматы, как pdf и djvu могут базироваться на изображении. В этом случае, стандартные методы получения текста применить будет невозможно).

После того, как текст был получен, он подвергается серии обработок. Во-первых, происходит удаление так называемых стоп-символов. Это слова, которые встречаются в каждом тексте и не несут в себе смысловой нагрузки (союзы, частицы, предлоги и другие). Во-вторых, над текстом выполняется операция стемминга. Стемминг – процесс нахождения основы слова. Его необходимо сделать для того, чтобы слова разных падежей, определялись как одно и то же слово. Для выполнения этой операции можно воспользоваться алгоритмом Портера [2].

Формирование частотного словаря слов

Построение частотного словаря слов – довольно простая задача, реализуемая огромным множеством известных алгоритмов. Однако, применительно к рассматриваемой задаче, недостаточно построения одного словаря слов.

Сначала текст разбивается на фрагменты определённой длины (длина – есть количество слов). Её значение должно быть достаточно большим, чтобы точность анализа была выше.

Для каждого фрагмента строится частотный словарь слов, а также производится оценка «дискриминантной силы» [3] каждого слова.

Анализ полученных данных

В конечном итоге смысл текста может быть выражен множеством наиболее значимых для него слов. Пересечение слов с самыми большими весовыми коэффициентами (дискриминантной силы) и слов с максимальной частотой позволяет сформировать конечный словарь.

Ориентируясь на закон Ципфа-Мандельброта [4], конечный словарь сортируется в порядке убывания значимости слов. Согласно принципу Парето 20/80 для повышения точности анализа необходимо отсечь последние 80% слов. Оставшиеся слова и будут являться смыслом анализируемого текста.

Взаимосвязь документов

Связывание документов будет производиться исходя из их полученной семантики – набора слов, отражающего смысл текста. В данном материале рассматривается Латентно-семантический анализ.

Латентно-семантический анализ – метод обработки текстов с целью их отображения на семантическое пространство – некое абстрактное представление семантической связности документов.

На первом шаге анализа требуется составить частотную матрицу индексируемых слов.

Следующим шагом эта матрица раскладывается на три составляющих по правилу сингулярного разложения [5]:

$$M = U \cdot W \cdot V^t$$

Матрица W является диагональной и содержит так называемые сингулярные коэффициенты, отсортированные в порядке убывания.

Далее для ускорения построения семантического пространства из матриц отбрасываются наименее значимые столбцы и строки соответственно, а также проводится нормализация.

Из полученных точек формируются группы любым известным алгоритмом кластеризации, например, методом k -средних.

Таким образом, в завершении работы алгоритма будет получен набор групп тематически и семантически связанных документов. Тематику каждой группы можно определить по словам, содержащимся в соответствующих кластерах.

Выводы

Хранилище данных, реализованное по представленным алгоритмам и принципам, без сомнения обладает рядом преимуществ. Это и возможность быстрого анализа информационных ресурсов, и группировка документов по их тематике, и обеспечение более точных результатов поиска.

Однако нельзя не отметить отрицательные стороны такого хранилища.

Во-первых, это сложность получения текста. Далеко не из всех документов можно выделить их текст. Очень часто документами являются фотографии и сканы других документов. В этом случае стоит использовать более сложные алгоритмы распознавания текста.

Во-вторых, время формирования такого хранилища данных гораздо более длительное, нежели обычной базы данных, основанной на реляционных принципах.

Таким образом, семантические хранилища информационных ресурсов представляют собой довольно удобный механизм для анализа хранимых документов. В зависимости от целей можно варьировать используемые алгоритмы для достижения лучшего результата. Но, как и любая другая технология, семантические хранилища информационных ресурсов следует использовать только для решения определённых задач.

Список используемых источников:

1. К. Маннинг, П. Рагхаван, Х. Шютце Введение в информационный поиск // М.: Вильямс, 2014.
2. M. F. Porter An algorithm for suffix stripping 1980.

3. Д. В. Ландэ *Оценки и визуализация уровня дискриминантной силы слов // Институт проблем регистрации информации НАН Украины.*
4. В. Дунаев *О ранговых распределениях в классификации // Научно-техническая информация. Серия 2, 1984.*
5. В. В. Стрижов *Информационное моделирование // Конспект лекций, сингулярное разложение.*

© 2016, Куликов Г.С., Курышов Д.М., Солнцев С.П.
Семантическое хранилище информационных
ресурсов

© 2016, Kulikov G.S., Kuryshov D.M., Solntsev S.P.
Semantic storage for information resources