Технология поиска SoftInform Search Technology

(c) Макс Магляс
SoftInform Inc.
max.maglyas@softinform.com

SoftInform Search Technology – технология поиска и обработки информации, содержащейся в текстовых файлах, базах данных и информационных системах. Она включает в себя все инструменты, необходимые для структуризации разрозненной информации в рамках предприятия и предоставляет собой эффективное решение любых проблем поиска и консолидации информации.

При поиске похожих по содержанию документов задействовано все множество слов встречающихся в документе с учетом всех словоформ и словаря синонимов. После обработки запроса в результирующем списке (с указанием процесса релевантности) выводятся документы максимально похожие на заданный фрагмент текста. 100% совпадения – найден документ-дубль. Документ же с меньшим процентом совпадения, соответственно, похож по содержанию на текст запроса. Следует отметить, что технология достаточно интеллектуальна для того, чтобы с высокой степенью точности определять релевантность искомого документа по отношению к запросу, не зависимо от изменений (удаление части текста, замена), внесенных в используемый в запросе текст.

Участие в РОМИП 2005

В конференции РОМИП компания СофтИнформ и поисковая система SearchInform участвовали впервые. Основная цель участия – проверка возможностей нашей поисковой системы и в частности, проверка работы заложенной в эту систему уникальной технологии поиска документов, похожих по содержанию.

Анализ результатов участия в конференции позволил нам определить как достоинства, так и некоторые недостатки разработанной компанией СофтИнформ технологии поиска. Участие в РОМИП 2005 оказалось для нас более чем плодотворным и мы сделали очень важные выводы, которые помогут нам в дальнейшей проработке и развитии нашей системы.

Фразовый поиск

Определение релевантности документа производится по нескольким параметрам:

- Совокупная длина фразы (СДФ) – это общая длина фразы в словах, то есть расстояние между первым и последним словом фразы.
- Количество слов из запроса, которые встретились во фразе (КНС) – количество слов, присутствующих во фразе. Обязательное условие этого параметра: между всеми словами фразы расстояние не может быть более заданного. В случае заданного расстояния более 100 слов поиск по фразе производиться не будет – в данном случае используется поиск по отдельным словам во фразе.

- Расстояние во фразе (РФ) – это максимально возможное из расстояний между словами во фразе.

Алгоритм определения релевантности:

1. Для оценки релевантности в первую очередь проводится анализ документов с максимальным количеством слов (КНС)
2. Следующим этапом производится оценка по расстоянию между словами во фразе (РФ). Чем расстояние меньше, тем документ, содержащий ключевую фразу запроса более релевантен.
3. Далее система сортирует найденные подходящие документы по СДФ. Документы с наименьшей совокупной длинной фразы считаются наиболее релевантными.

Поиск похожих

Для определения похожести документов и их релевантности используется 4 коэффициента, взаимодействующих между собой: при расчете общего «веса» документа каждый из коэффициентов влияет на остальные.

- Статистика: количество совпадающих с запросом слов в анализируемом документе. При этом учитывается вес слов (более редкие имеют больший вес). Данный коэффициент в общей составляющей играет минимальную роль. Он скорее корректирующий, а не основной в процессе определения релевантности.
- Коэффициент фраз: наиболее релевантным (похожим по содержанию) документом считается тот, в котором с текстом запроса наиболее точно совпадают маски фраз. Система производит «нечеткий» анализ – рассматривается наличие посторонних или пропущенных в той или иной фразе слов по сравнению с запросом.
- Коэффициент «СофтИнформ 1»: собственная запатентованная разработка компании СофтИнформ. Не разглашается.
- Коэффициент «СофтИнформ 2»: собственная запатентованная разработка компании СофтИнформ. Не разглашается.

После расчета всех коэффициентов релевантности найденных документов выводится единое число «похожести». Назовем его «S». Точно таким же вышеописанным образом рассчитывается общее значение «документа-запроса» - «К». Собственно, коэффициент похожести того или иного документа на текст запроса вычисляется путем сравнения значений «запроса» и найденного документа.

Возможности SoftInform Search Technology

- Поиск документов похожих по содержанию на текст запроса (сокращение времени на подбор ключевых фраз и на просмотр ненужных документов)
- Решение проблемы размытость информационного наполнения (использование в работе одного нужного документа, а не нескольких его дублей или похожих на него, но из других источников)

- Консолидация информации из различных источников (поиск и обработка информации из различных баз данных, информационных систем и так далее)

- Обработка и создание отчета похожести документов, уже находящихся в базе (выявление дублей)

Высокая скорость индексирования (до 6 Гб/час), малый размер индекса (15-20% от реального объема текстовой информации), поддержка практически всех распространенных форматов текстовых файлов (включая .pdf и .html) и корректная работа с архивами делают SoftInform Search Technology незаменимым инструментом поиска информации.

Универсальные источники данных

Технология прекрасно работает с наиболее распространенными форматами текстовых файлов (txt, doc, rtf, pdf, htm, html), поддерживая и корректно обрабатывая все из них. Но в крупных организациях, где информация обычно содержится в различных информационных системах -CRM, архивах, СУБД и так далее этого недостаточно. Технология поиска похожих справляется и с этой задачей. В ней встроена возможность индексации полей из практически всех существующих на данный момент распространенных систем (например, Access, MS SQL, Oracle, а также любых СУБД, поддерживающих SQL).

Также не составляет никакого труда адаптировать технологию (при внесении минимальных корректировок) под любую другую базу данных или информационную систему. Причем, источники данных, доступные для индексации нашей программой могут быть различны и могут находиться в разных местах.

Заключение

Технология поиска похожих документов «СофтИнформ» – это незаменимый инструмент для экономии времени и денег, позволяющий любой компании резко сократить затраты (как материальные, так и временные) на поиск и обработку информации в больших объемах данных. Технология поиска похожих документов «СофтИнформ» это:

- быстрый и точный поиск похожих по содержанию документов в любых объемах данных.

- возможность исключить дублирование информации

- интеграция в любые базы данных и прикладные системы, а также работа с любыми документами

- возможность разработки широкого спектра приложений как для локальных внутрикорпоративных сетей, так и для глобальных интернет-решений