Поисковая система “mnoGoSearch”

© Барков А.И., Барков И.А.

Lavtech.Com. Corp.
bar@mnogosearch.org

Аннотация

Настоящая работа является отчетом об участии в конференции РОМИП-2004. Главной целью работы была апробация частотных методов оценки релевантности документа запросу при поиске по Web-страницам.

Введение

MnoGoSearch является свободно распространяемой поисковой системой, работающей в операционных системах семейства Unix, предназначенной для организации поиска на одном или многих Web-серверах. Первая версия mnoGoSearch была выпущена в ноябре 1998 под названием UDMSearch. В октябре 2000 года проект был передан Lavtech.Com.Corp. В октябре 2001 года появились коммерческие модификации системы, реализованные для операционных систем Windows. Последние версии системы можно найти на сайте http://www.mnogosearch.org/.

1. Краткое описание системы

mnoGoSearch состоит из двух частей. Первая часть - индексирующий механизм (**indexer**). **Indexer** пробегает по ссылкам и сохраняет в базе данных информацию о документах, терминах и ссылках. Вторая часть состоит из CGI-программы, предоставляющей возможность поиска в данных, собранных **indexer’**ом. Особенностью системы является то, что она использует в качестве хранилища базу данных, поддерживающую язык SQL. Система протестирована с наиболее популярными в настоящее время СУБД (около 10-ти).

Основные возможности mnoGoSearch включают:

- Поддержку основных протоколов Интернета (HTTP, HTTPS, FTP, NNTP);

- Встроенную поддержку документов TXT, HTML, XML, а так же возможность подключения внешних программ-конверторов для любых других типов документов, таких как DOC, PDF, RTF, XLS, PPT;
- Нечёткий поиск на основе синонимов, подстрок, а так же генерации словоформ (падежи, склонения, и т.д.);
- Мощную языковую поддержку, позволяющую организовать поиск по документам на многих языках мира, включая языки Восточной Азии. В состав системы входит распознаватель языка документа;

2. Направления развития системы

С самого начала mnoGoSearch позиционировалась для работы с небольшими объемами данных. Однако последние версии системы сделали возможным индексировать до нескольких сотен тысяч документов. В ближайших версиях системы планируется добавление возможности масштабирования на нескольких компьютерах. В связи с этим проблема ранжирования выдаваемых на запрос документов стала более актуальной.

3. Опыт участия в РОМИП 2004

В конференции РОМИП мы участвовали впервые. Участие в конференции совпало с введением в mnoGoSearch алгоритмов вычисления релевантности документов. Нами было проведено сравнительное исследование пяти классических моделей, рассмотренных далее.

В описании моделей используются следующие обозначения:

N – число документов в наборе документов;

M – число терминов в наборе документов;

D_j – вектор терминов, присвоенных документу j;

d_{ij} – вес термина i, в документе j;

$(TF)_i$ – (Term Frequencies) суммарная частота появления термина i в наборе документов.

Частотная модель 1 определяет взвешивающую функцию значимости термина i как

$$W_i^{'} = (TF)_i / N .$$

Частотная модель 2 имеет вид:

$$W_i^{''} = (TF)_i \, (IDF)_i .$$

Модель 3 учитывает различительную силу термина и имеет вид:

$$W_i''' = (TF)_i (DV)_i,$$

где W_i''' есть вес, присвоенный термину i, а $(DV)_i$ - значение его различительной силы.

Модель 4 "сигнал – шум". Шум термина будем определять следующим образом:

$$N_i = \sum_{j=1}^{N} \frac{d_{ij}}{(TF)_i} log \frac{(TF)_i}{d_{ij}}.$$

Сигнал будем определять формулой:

$$S_i = log(TF)_i - N_i.$$

Значимость термина i в этом случае будет

$$W_i'''' = \frac{S_i}{N_i}.$$

Модель 5 известна как модель среднеквадратичного отклонения. Если

$$\overline{c_i} = \frac{\sum_{j=1}^{N} d_{ij}}{N} -$$

средняя частота термина в документе, то среднеквадратичное отклонение равно

$$(V_i)^2 = \frac{\sum_{j=1}^{N} (d_{ij} - \overline{c_i})}{N-1}.$$

Тогда значимость термина i в этом случае определяется как

$$W_i''''' = \frac{(TF)_i \cdot (V_i)^2}{(\overline{c_i})^2}.$$

Поведение всех пяти моделей было промоделировано в системе Excel. Мы не стали вводить какие-то точные методы оценки моделей, и сделали выбор интуитивно. На разных наборах документов и терминов все модели вели себя по-разному. Однако, по нашему мнению, модели 2, 3, 5 выглядели значительно лучше, чем модели 1 и 4. Окончательный выбор пал в пользу метода

среднеквадратичного отклонения, который и был применен для прогона конкурсного задания.

Справедливости ради нужно отметить, что мы ставили перед собой цель исследовать частотные методы в чистом виде. Поэтому, в качестве веса термина в документе использовалось количество появлений слова в документе, а такие факторы, как близость слов запроса друг к другу и наличие слов в заголовке документов не были учтены, что, безусловно, сказалось на результатах оценки.

Предварительный анализ результатов нашего участия в РОМИП-2004 хорошо позволил нам увидеть как достоинства, так и недостатки разрабатываемой поисковой системы, что неоценимо для правильного выбора направлений дальнейшей работы. Поэтому считаем, что участие в конференции оказалось для нас плодотворным.

В качестве замечания отметим, что информации, опубликованной на сайте РОМИП, не всегда было достаточно. Мы благодарны Игорю Некрестьянову за постоянную поддержку, без которой наше участие в конференции значительно бы усложнилось.

Search engine “mnoGoSearch”

Barkov A.I. Barkov I.A.

This article presents a report on experiments in web page retrieval that were made as a part of RIRES initiative. The main goal of these experiments was to approbate frequency-based relevancy calculation in Web search.