SSSleuth на РОМИП 2008

© Сергей Крылов

Kryloff Technologies
http://www.kryltech.com/feedback.htm

Аннотация

Настоящая статья посвящена апробации поисковой системы SSSleuth™, разработанной автором. Указаны уникальные некоторые из особенностей алгоритмов системы, а также цель участия в семинаре.

1. Введение

Kryloff Technologies первый раз принимает участие в семинаре РОМИП. На РОМИП 2008 автор ставил перед собой задачу выяснить, насколько качество поиска системы SSSleuth, реализованной в многоязычном варианте, отличается от систем, разработанных для поиска преимущественно на русском языке, а также найти партнёров по дальнейшим совместным исследованиям и разработкам.

2. Поисковый алгоритм системы SSSleuth

Так же, как и другие подобные системы, SSSleuth используется для задач поиска релевантных документов по запросу пользователя и ранжирования отобранных документов по степени соответствия запросу. Кроме решения задачи поиска, система SSSleuth также включает в себя модуль контекстного аннотирования для выделения найденных системой фрагментов исходных документов в отчётах системы, а также модуль построения резюме и автоматического определения языка текстов документов. Однако, поскольку на данный момент автор принял участие только в дорожках, связанных с классической задачей поиска, далее будут указаны особенности только поисковой части системы.

2.1 История появления алгоритма SSSleuth

Появление системы SSSleuth явилось результатом глубоких математических исследований, проводимых автором начиная с 1992 года. В результате исследований автору удалось открыть фундаментальный и универсальный метод поиска, аналоги которому в мире автору неизвестны, а также создать ряд программных продуктов, распространяемых компанией Kryloff Technologies по всему миру.

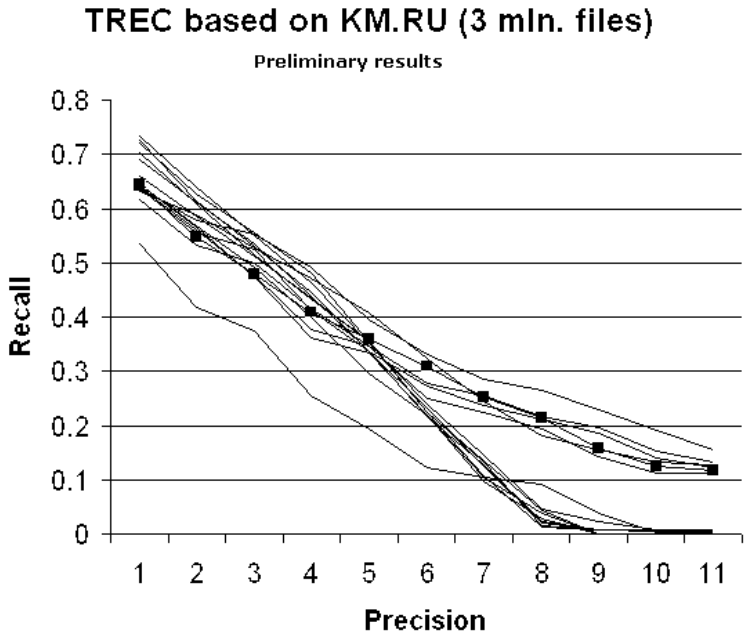
2.2 Уникальные особенности поискового алгоритма SSSleuth

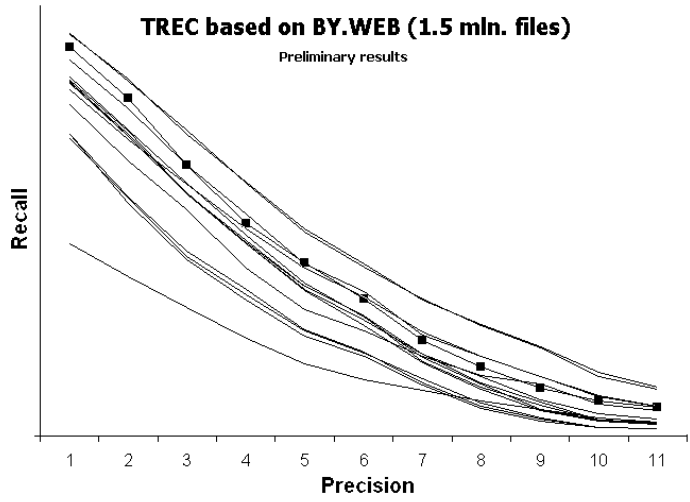
2.2.1. Многоязыковая поддержка.

На данный момент и в той реализации, что применялась для построения отчётов в тестовых дорожках РОМИП, система SSSleuth производит поиск на сорока языках, включая большинство европейских языков, таких как русский и английский, а также арабский, индонезийский, иврит, корейский, китайский и японский. При этом алгоритм поиска остаётся неизменным на всех из перечисленных выше языков. Более того, включение в систему возможности поиска на дополнительных языках автоматизировано и, по существу, требует лишь указания системе алфавитов этих языков.

2.2.2. Простота системы.

Объём исходных текстов существующих на данный момент на рынке поисковых систем, по утверждению их разработчиков, составляет десятки, а иногда и сотни мегабайтов. Объём же исходных текстов SSSleuth составляет всего несколько десятков килобайтов, из чего следует, что SSSleuth является существенно, в тысячи раз более простой системой. Только непосредственно в процессе кодирования таких поисковых систем, как “Google” и “Яндекс”, участвуют десятки программистов; SSSleuth же полностью разрабатывался одним человеком (информация об объёмах текстов и количестве разработчиков получена с официальных сайтов компаний “Google” и “Яндекс”). Однако, несмотря на эти факты, SSSleuth показывает достаточно высокое качество поиска, которое находится на уровне лучших из систем, принимаюших участие в семинаре РОМИП. Ниже приведёны 11-точечные графики TREC, построенные на основании ответов систем по коллекциям KM.RU и BY.WEB; чёрными квадратиками выделены графики SSSleuth:


TREC based on KM.RU (3 mln. files)
Preliminary results
Recall
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1 2 3 4 5 6 7 8 9 10 11
Precision


TREC based on BY.WEB (1.5 mln. files)
Preliminary results
Recall
1 2 3 4 5 6 7 8 9 10 11
Precision

2.2.3. Способность находить приближённые соответствия.

В той реализации SSSleuth, что принимала участие в семинаре РОМИП, в качестве или, точнее, вместо блока, который бы выделял слова или термы как единицы поиска, использовались трёхбуквенные Q-Термы: каждое слово из индексируемых документов разбивалось на последовательности из трёх следующих друг за другом символов алфавита языка, на котором ведётся поиск (русского в данном случае). После индексации Q-Термов, система выделяла документы, в которых большинство или все Q-Термы, на которые разбивается текст запроса, встречались в исходных документах примерно в той же последовательности, что и в тексте запроса.

Например, фамилия “Домогаров”, участвовавшая в одном из запросов по коллекциям KM.RU и BY.WEB, разбивается следующим образом: <дом><омо><мог><ога>…<ров>; в угловых скобках заключены единицы поиска, которые SSSleuth рассматривает как неделимые элементарные элементы. Базовый алгоритм поиска SSSleuth легко находит “домОгаров” при запросе “домАгарова” или “домАгаров” (буква “А” вместо ”О”), поскольку общими в запросе и в тексте документа для обоих слов являются Q-Термы <дом>,<гар>,<аро>,<ров>, причём идущие в той же самой последовательности. То же верно и для словосочетаний, в которых присутствуют слова, не участвующие в тексте запроса: часть слов могут быть изменены либо отсутствовать вовсе.

В качестве развёрнутого примера автор приглашает посетить Web-страницу [2], в которой производится сравнение SSSleuth с поисковой системой Google Desktop на английском языке, причём на естественных языковых запросах (система успешно отвечает на такие запросы, как “какой самый большой телескоп в мире?”, “почему луна постоянно повёрнута к нам одной стороной?”, и т.п.). Там же Вы можете увидеть примеры того, как SSSleuth производит контекстное аннотирование в своих отчётах.

2.2.4. Работа с огромным количеством элементарных единиц поиска.

Хотелось бы обратить внимание на то, что количество Q-Термов, выделяемых SSSleuth из документа, приблизительно равно длине текстовой части документа в байтах. Это на порядок превосходит

количество единиц поиска для систем, основанных на словах как неделимых единицах поиска. Таким образом, если корпус состоит из 3-х миллионов документов (как, например, KM.RU), то SSSleuth фактически производит поиск в 30-ти миллионах. Как следует из приведённого выше графика TREC, система вполне успешно справляется со сложной задачей обнаружения последовательностей термов, которые группируются в текстах документов, причём в последовательности, максимально приближенной к образцу.

Несмотря на универсальность и много-язычность подхода, связанного с Q-Термами, автор уверен, что наличие более продвинутого блока выделения элементарных элементов поиска, ориентированного на конкретный язык, способно существенно:

a) повысить и так уже достаточно высокое качество поиска системы, а также
b) сократить объём индексных файлов и
c) ещё более ускорить процесс поиска.

Автор ставил такие эксперименты по запросам на английском языке, для которого у автора имеется простой стеммер (блок выделения корневых составляющих слов, которые в дальнейшем рассматриваются SSSleuth как элементарные единицы поиска вместо трёхбуквенных Q-Термов). Результаты экспериментов показали большую перспективность этого направления, однако обширные эксперименты на русском языке автором до настоящего момента не проводились.

3. Заключение.

По сути, с тем простым блоком “синтаксического разбора”, с которым SSSleuth участвовал в семинаре, система ведёт поиск всего лишь на фонетическом уровне, показывая при этом, однако, вполне приемлемое (и ожидаемое автором) качество. Этот факт связан с тем, что базовый алгоритм поиска SSSleuth, открытый автором в результате его многолетних математических исследований, не требуя перебора вариантов, связанных с поиском близлежащих слов или термов, выделяемых системой в процессе индексации, всё же допускает существенные отличия от задаваемого образца поиска: множество примеров (на английском языке) дано в [2].

Как уже было сказано выше, основной целью участия в семинаре автор видит для себя поиск партнёров, с которыми он хотел бы продолжить исследования, направленные на создание существенно более продвинутых систем поиска с намного более высокими характеристиками, чем те, что участвуют в семинаре, включая SSSleuth в его многоязычной реализации.

Автор уверен, что даже простая замена Q-термов, выделяемых системой и используемых ею в данный момент в качестве элементов поиска, на канонические формы слов (например, русского языка) уже способна вывести систему в лидеры по качеству поиска не только в семинаре РОМИП, но и во всём мире – такова сила и мощь базовых универсальных алгоритмов поиска, изобретённых автором и используемых им в SSSleuth. Другие приоритетные направления возможных совместных с партнёрами исследований включают:

a) оптимизация параметров базовых алгоритмов SSSleuth;
b) их адаптация для конкретного языка (русского, английского, арабского, и т.д.);
c) включение в процесс поиска этапа, связанного с формулировкой и последующим нахождением одного или серии образцов поиска на языке запроса. Учитывая эффективность неперебороного алгоритма SSSleuth, направленного прежде всего на поиск термов, не просто присутствующих в документах, но и находящихся в той же последовательности, что и в образце, с помощью SSSleuth становится возможным быстрый поиск сразу серии образцов. Образцы следует формулировать на основании текста запроса, заменяя его на синонимичные конструкции, ожидаемые в текстах документов. Последующий поиск таких конструкций наряду с исходным текстом запроса должен, по мнению автора, ещё более повысить качество поиска и поэтому рассматривается им как одно из приоритетных направлений.

Автор просит считать данную работу официальным приглашением заинтересованных сторон к дальнейшим экспериментам и плодотворному взаимному сотрудничеству.

Литература

[1] ROMIP Web site, 2008. http://www.romip.ru
[2] S. Kryloff. SSSleuth and Google Desktop: Comparison Results, 2008. http://www2.kryltech.com/google/google.htm
[3] Kryloff Technologies Web site, 2008. http://www.kryltech.com

SSSleuth at RIRES-2008

© Sergey Kryloff

Kryloff Technologies

http://www.kryltech.com/feedback.htm

The paper presents information retrieval system SSSleuth developed by Sergey Kryloff for RIRES-2008; the system participated in ad hoc tracks on various collections. This article describes distinctive and unique features of SSSleuth methods, experimental results, and finally, invites all concerned parties to run further joint development and investigations, which would be based on mutual achievements in the Full-Text Search and Retrieval area.