ЛОГИКА БИЗНЕСА 2.0
МИССИЯ ВЫПОЛНИМА

РАЗВИТИЕ МОДЕЛИ, ОСНОВАННОЙ НА ЗНАНИИ ОБ АВТОРАХ, ДЛЯ ПОИСКОВЫХ ПРИМЕНЕНИЙ

Валентин Молоканов, Дмитрий Романов, Валентин Цибульский

Центр компетенций интеллектуальных поисковых систем и текстовой аналитики

Диалог 2013, 1 июня 2013 г.

О настоящей презентации

Презентация состоит из следующих основных разделов:

- Принципы предлагаемой технологии (поискового ядра)
- Поиск экспертов как обучающая задача на серии результатов конференций Text Retrieval Conference (TREC)
- Результаты прогона заданий TREC 2006-2007
- Планируемая сфера применения
- Выводы

Модель поиска экспертов

Основная идея

- Найти людей, которые смогут решать задачи бизнеса
- Человеко-ориентированный подход позволяет избавиться от перегрузки индекса большими объемами данных, а также от сохранения полных исходных сообщений, противоречащего корпоративным стандартам
- Информация об авторстве документов позволяет моделировать сеть коммуникаций между людьми в коллекции
- Лексику и авторов в коллекции можно различать с помощью специально выбранных метрик: значимость и сила связи

Модель поиска экспертов

Значимость лексики

Частота употребления термина автором

$$f(t, p) = \frac{n(t, p)}{N(p)}$$

Логарифм частоты, усредненный по авторам

$$e(t) = \frac{1}{P(t)} \sum_p \ln f(t, p)$$

Дисперсия данного распределения

$$D(t) = \left[\sum_p (\ln f(t, p) - e(t))^2 \right]^{1/2}$$

Значимость термина

$$S(t) = -\frac{D(t)e(t)}{P(t)}$$

Модель поиска экспертов

Сила связи лексики с автором

$$L(t,\ p) = \ln(n_s(t,\ p)+1) + \ln(n_r(t,\ p)+1) + \ln(c_s(t,\ p)+1) + \ln(c_r(t,\ p)+1)$$

Запрос

Автор

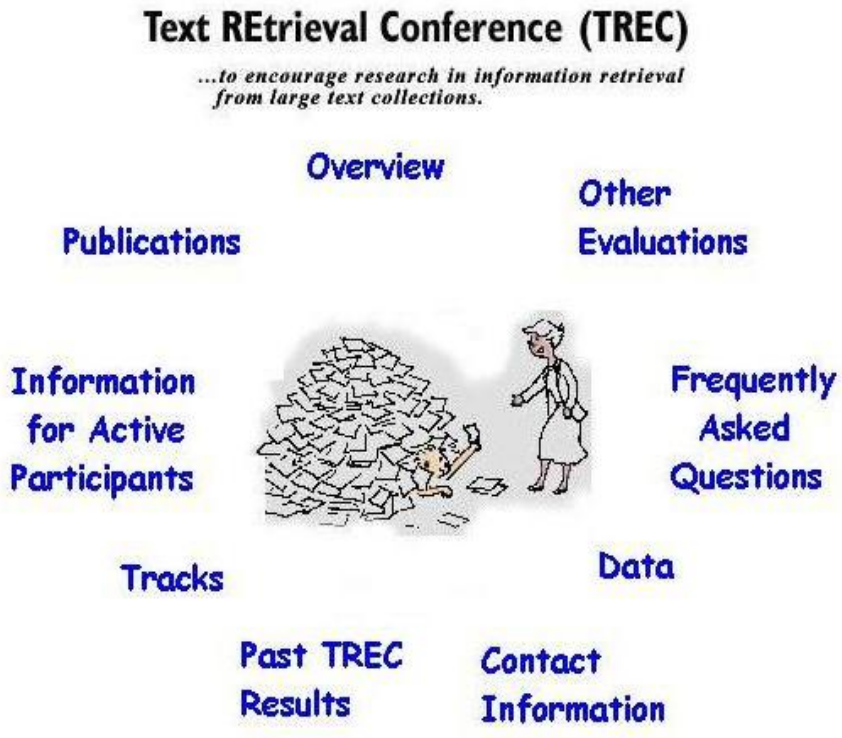
Модель поиска экспертов

Функция ранжирования

$$W(p) = \sum_{i} \sum_{x_i \in X_i} C_i \, S(x_i) \, L(x_i, \, p)$$

Text Retrieval Conference (TREC)

- TREC – серия ежегодных конференций, сосредоточенных на исследовании различных областей информационного поиска и предусматривающих для каждой области соответствующую «дорожку»
- Поиск экспертов существовал в 2005-2008 гг. в виде отдельного задания на корпоративной дорожке (Enterprise track)
- Для оценки производительности движка на TREC приняты специальные метрики эффективности поиска. Это макроусредненная средняя точность (MAP) и точность на 5-м (P@5) и 20-м (P@20) уровнях.



Результаты

5 лучших автоматических пусков TREC

TREC 2006

Run	MAP	P@5	P@20
hse2006qMod	**0.5954**	**0.6200**	**0.5130**
SJTU04	0.5947	0.8245	0.6031
PRISEXB	0.5564	0.7592	0.5459
UMaTDFb	0.5016	0.7265	0.5000
THUPDDSNEMS	0.4954	0.6694	0.5071

TREC 2007

Run	MAP	P@5	P@20
THUIRMPDD4	0.4632	0.2280	0.0910
SJTUEntES03	0.4427	0.2360	0.0910
ouExTitle	0.4337	0.2520	0.0950
hse2007Ent	**0.3930**	**0.2120**	**0.0840**
ExpertRun02	0.3689	0.2040	0.0790

Сфера применения

- поиск экспертов
- классификация информации (плоская, иерархическая, многомерная)
- правовая экспертиза
- поиск заимствований

Выводы

- Вычисление силы связи автора с лексикой позволяет выявить определенные термины, которые являются характеристическими для автора и с помощью которых можно его находить
- Взвешивание лексики дает возможность выделить из всей коллекции небольшую порцию лексики – значимую лексику; именно она является необходимой для точного ответа на запрос
- Полученные результаты превосходят по точности большинство результатов TREC, при этом не используются ни дополнительные признаки в информации (например, структура документов), ни информация извне коллекции
- Поисковое ядро универсально и не зависит от коллекции
- Помимо высокой эффективности, поисковое ядро предоставляет возможности для дальнейшей оптимизации, а также для расширения сферы применения

Спасибо за внимание!