Mail.Ru на РОМИП 2007

Костин Михаил, Проскурин Андрей

Mail.Ru

kostin@corp.mail.ru, proskurin@corp.mail.ru

Аннотация

Статья посвящена участию компании Mail.Ru в семинаре РОМИП-2007. Основное внимание уделено экспериментам с различной трактовкой значимых зон html документа в алгоритмах учета близости расположения слов запроса в документе.

1. Введение

В 2007 году мы в третий раз приняли участие в семинаре РОМИП. Участвуя в поисковых дорожках семинара, мы ставили перед собой цель проверить некоторые гипотезы об особой трактовке значимых зон документа (в первую очередь, заголовков) в алгоритмах учета близости слов.

2. Основные принципы ранжирования

В качестве базы для своих экспериментов мы использовали алгоритм ранжирования уже проходивший проверку на семинарах РОМИП прошлых лет [1] (с некоторыми модификациями, не имеющими в данном случае существенного значения). Здесь мы отметим только те его особенности, которые необходимы для понимания сути проведенных нами экспериментов.

2.1 Функция релевантности

Базовая функция релевантности в нашем алгоритме ранжирования выглядит следующим образом:

$$W = K_f W_f + K_p W_p + K_{ps} W_{ps} \quad (1),$$

где Wf - частотность термов запроса по TF*IDF,

Wp - встречаемость пар соседних слов запроса,

Wps - вес самого релевантного пассажа в документе,

Kf, Kp, Kps - коэффициенты.

2.2 Пассажи

Наши эксперименты в этом году касались последнего слагаемого в формуле 1 – веса лучшего из релевантных пассажей в документе. Под релевантным пассажем мы понимаем фрагмент текста документа, не превышающий заданного ограничения по длине и содержащий представительное множество слов запроса. Принадлежность слов к различным зонам (html-тегам) документа (title, bold, h1-h6, обычный текст) в нашем базовом алгоритме учитывается при вычислении веса пассажей, однако не влияет на алгоритм выделения пассажей в тексте.

3. Пассажи с «джокерами»

Основную идею, подвергнувшуюся проверке в наших экспериментах, можно сформулировать следующим образом: слова, расположенные в наиболее значимых частях документа (прежде всего, в заголовке) можно считать расположенными рядом с любым словом документа, так как они отражают смысл документа в целом. То есть, в алгоритме поиска пассажей, слова, встретившиеся в этих частях документа, рассматриваются как своеобразные «джокеры», которые можно подставить в любое место документа для получения релевантного пассажа.

Приведем пример. Запрос пользователя: «Значение имени Анастасия». Заголовок документа: «Значения женских имен». Фрагмент текста: … Анастасия – воскресшая (греч.)…

Очевидно, что документ релевантен запросу, однако при ранжировании по нашему базовому алгоритму он может получить достаточно небольшой вес ввиду отсутствия в нем пассажей, содержащих все слова запроса. В то время как при поиске пассажей по алгоритму, учитывающему слова заголовка указанным выше способом, такой пассаж будет выделен и документ получит адекватный своей релевантности вес.

Конечно, так бывает не всегда. Можно с легкостью привести противоположные примеры, когда такой учет слов заголовка напротив, необоснованно завышает релевантность документа. Поэтому определить жизнеспособность данного алгоритма невозможно без практической проверки.

4. Эксперименты РОМИП

Нами было предоставлено три прогона по веб-коллекции и два прогона по коллекции нормативно правовых документов.

Прогоны по веб-коллекции:

- Только обычные пассажи (Standart)
- Пассажи с «джокерами» из заголовков (Title)
- Пассажи с «джокерами» из заголовков, тегов h1-h3 и начала текста документа (первые 100 слов) (Headers)

Прогоны по коллекции нормативно-правовых документов:

- Только обычные пассажи (Standart)
- Пассажи с «джокерами» из заголовков (Title)

Пассажам с «джокерами» (если они учитывались) нами давался несколько меньший вес, чем обычным пассажам, в силу меньшей степени их достоверности.

5. Результаты

В таблице приведены результаты наших прогонов по веб-коллекции (объединение коллекций km.ru и белорусского интернета) для сильных и слабых требований к релевантности [2].

Прогон	Web adhoc, pd 50, OR		Web adhoc, pd 50, AND	
	Precision 10	Average Precision	Precision 10	Average Precision
Standart	0.433	0.249	0.267	0.246
Title	0.415	0.242	0.26	0.244
Headers	0.432	0.252	0.252	0.241

Таблица 1. Результаты прогонов по Веб-коллекции.

Во второй таблице приведены результаты наших прогонов по коллекции нормативно-правовых документов.

Прогон	Legal adhoc, pd 50, OR		Legal adhoc, pd 50, AND	
	Precision (10)	Average Precision	Precision 10	Average Precision
Standart	0.822	0.418	0.586	0.418
Title	0.824	0.422	0.588	0.422

Таблица 2. Результаты прогонов по коллекции нормативно-правовых документов.

Из полученных результатов можно видеть, что:

- использование алгоритмов с джокерами не оказало большого влияния на результаты, разница между оценками релевантности по разным прогонам достаточно невелика;
- результаты оказались достаточно противоречивыми, по разным критериям оценки прогоны располагаются в разном порядке.

Причина этого, вероятно, в первую очередь заключается в слишком небольшом объеме экспериментального материала для оценки (тут следует учесть, что далеко не для всех запросов вообще нашлись документы, чей вес в разных прогонах отличается).

Некоторой неожиданностью для нас оказались результаты третьего прогона по веб-коллекции (в котором близкими к любому слову документа считались слова не только из заголовка, но и из тегов h1-h3 и начала текста). Использованный там алгоритм очень грубый (например, содержание тегов h1-h3, как правило, описывает только ту часть документа, к которой они, относятся, а не весь документ в целом) и мы ожидали некоторого падения релевантности на нем, которого, однако, не наблюдается. Впрочем, это тоже можно объяснить недостаточным объемом экспериментальных данных.

В этом году сравнение эффективности наших алгоритмов с алгоритмами других участников семинара не было нашей главной целью, тем не менее, такое сравнение всегда интересно. К сожалению, сравнительные результаты дорожки поиска по веб-коллекции оказалось сложно правильно интерпретировать в силу того что нами были предоставлены результаты прогонов по суммарной коллекции (km.ru и белорусский интернет), а другими участниками по каждой из двух коллекций в отдельности (или по одной из них). Приведем сравнительные результаты дорожки поиска по нормативно-правовой коллекции.

	Run 1	Run 2	Mail.Ru (Standart)	Mail.Ru (Titles)	Run 5
Precision (10)	0.45	0.78	0.822	0.824	0.580
Average Precision	0.131	0.391	0.418	0.422	0.231

Таблица 3. Результаты участников дорожки поиска по коллекции нормативно-правовых документов (pd 50, OR).

6. Заключение

В целом, можно сделать вывод, что гипотезы, проверке которых было посвящено наше участие в РОМИП 2007, требуют дальнейшего изучения. Можно предположить, что положительного результата можно достичь на пути более гибкого использования изложенного метода, учитывающего особенности конкретного запроса и конкретного документа.

Литература

[1] А.Федоровский, М. Костин, А. Проскурин. Mail.Ru на РОМИП-2005. Труды третьего российского семинара по оценке методов информационного поиска. Под ред. И.С. Некрестьянова - Санкт-Петербург: НИИ Химии СПбГУ, 2005.

[2] Труды РОМИП'2006, Под ред. И.С. Некрестьянова - Санкт-Петербург: НУ ЦСИ, 2006