KM.RU на РОМИП-2007

Сергей Татевосян,
Наталья Брызгалова
«КМ онлайн»

KM.RU на РОМИП-2007:

- Представление новой поисковой системы, разработанной KM.RU
- Тестирование алгоритма, участие в дорожке поиска web-adhoc

Алгоритм поиска и ранжирования:

- ***Текст запроса***: все слова из запроса - ключевые
- ***Веб-документ***: вес данного документа в коллекции
- ***Текст документа***: присутствие в документе ключевых слов; расстояние между ключевыми словами в документе
- ***Html-разметка***: важны элементы разметки, которые выделяют значимые части документа
- ***Гиперссылки***: наличие гиперссылок с других документов на данный

Общая формула релевантности

$$W = W1 + W2 + W3,$$

где *W* – итоговое значение релевантности документа

W1 **- информационная значимость документа и его вес в коллекции**

W2 **- информационная значимость ссылок на документ**

W3 **- учет расстояния между словами запроса в документе**

W1 - информационная значимость документа

W1 = TF*IDF(Doc) * F1(DocWeight)

- *TF*IDF(Doc)* вычисляется по модификации BM25
- *F1(DocWeight)* – функция от веса документа. Особенности:
  - приведение значения DocWeight до нужного диапазона
  - часть ссылок признаются неинформативными и в расчете не участвуют

W2 – информационная значимость ссылок

W2 = Σ (TF*IDF(Link)* F2(LinkWeight))

- *TF*IDF(Link)* - *TF*IDF* ссылки на данный документ
- *F2(LinkWeight))* – функция приведения весов ссылок на документ. *LinkWeight* вычисляется аналогично *DocWeight*

W3 – учет расстояния между словами

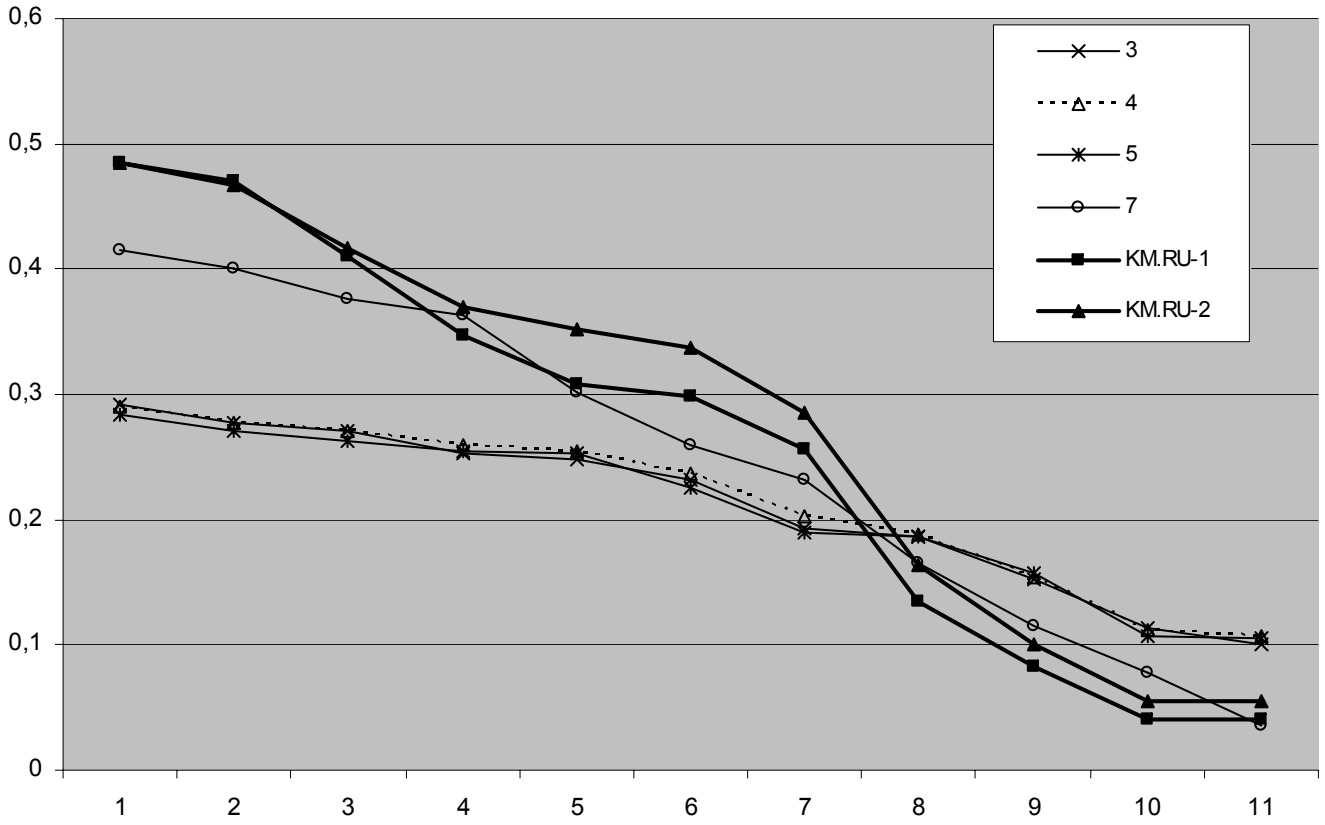
W3 = F3(paccm)

Имеет ненулевое значение при прохождении кворума, вычисляющегося по сумме IDF слов

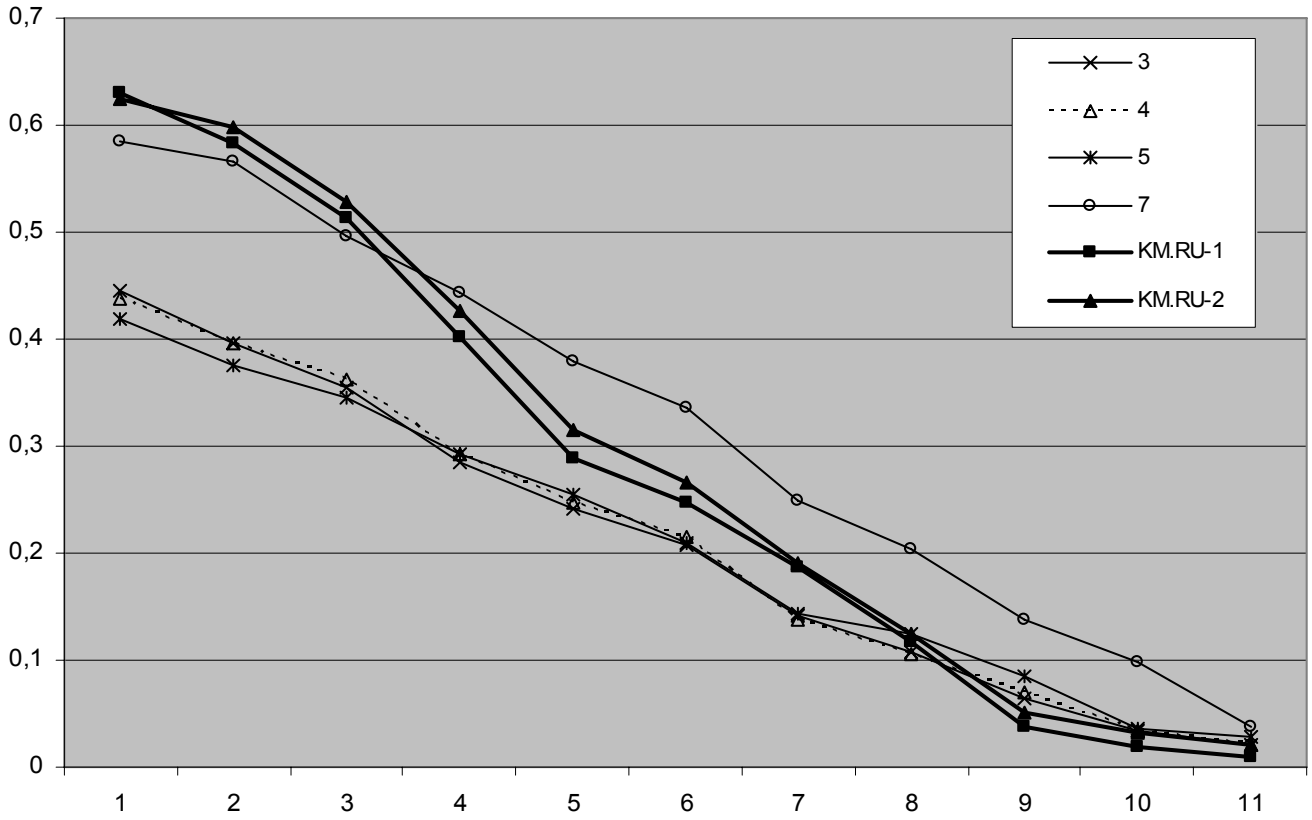
Прогоны

- ***Коллекции***
  - KM.RU: уже тренировались на этой коллекции; прогоны осуществляли по частично очищенной от дублей коллекции (из 3 млн. документов осталось 850 тыс.)
  - BY.WEB: новая для нас коллекция
- ***Модификации алгоритма***
  - Прогон 1: *W3 = 0*
  - Прогон 2: *W3 = F3(paccm)*

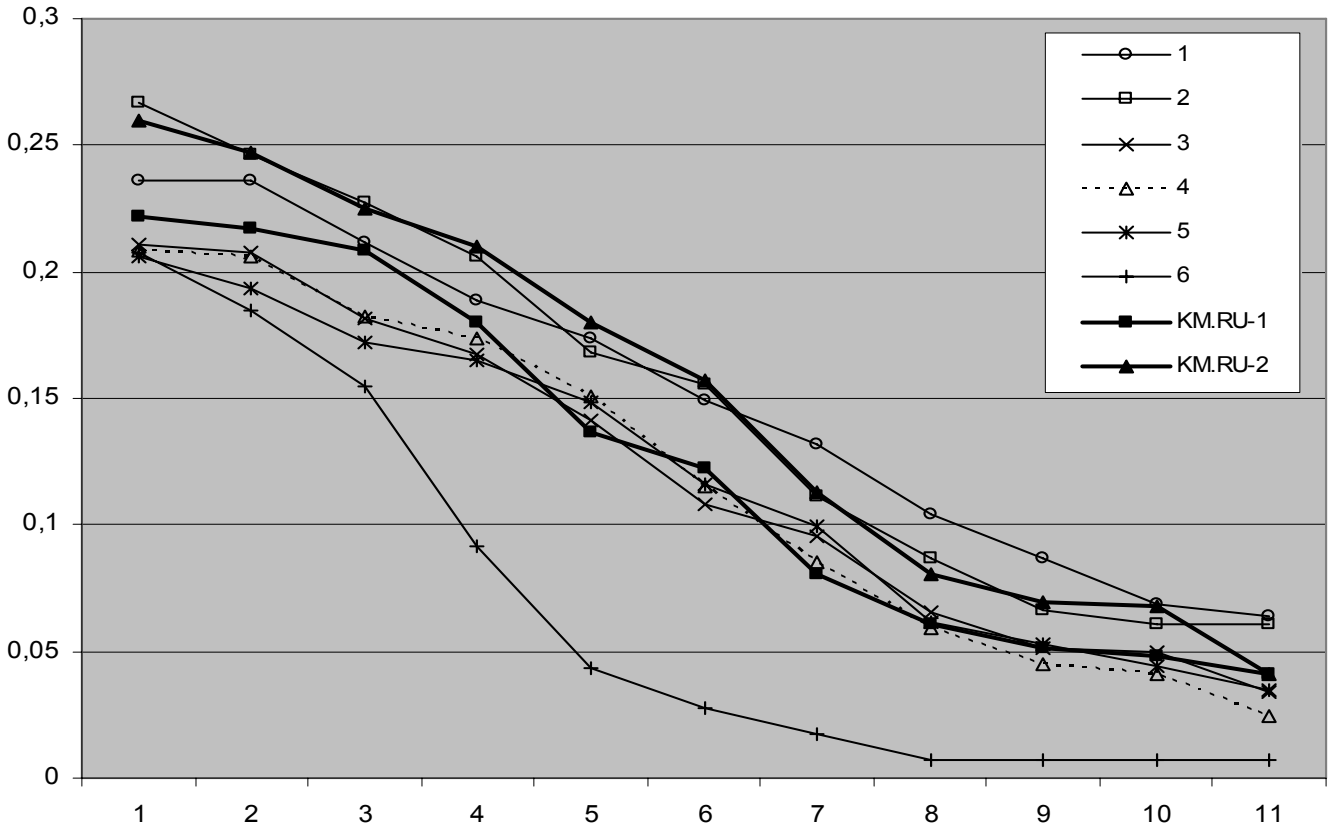
Web adhoc, km.ru, AND



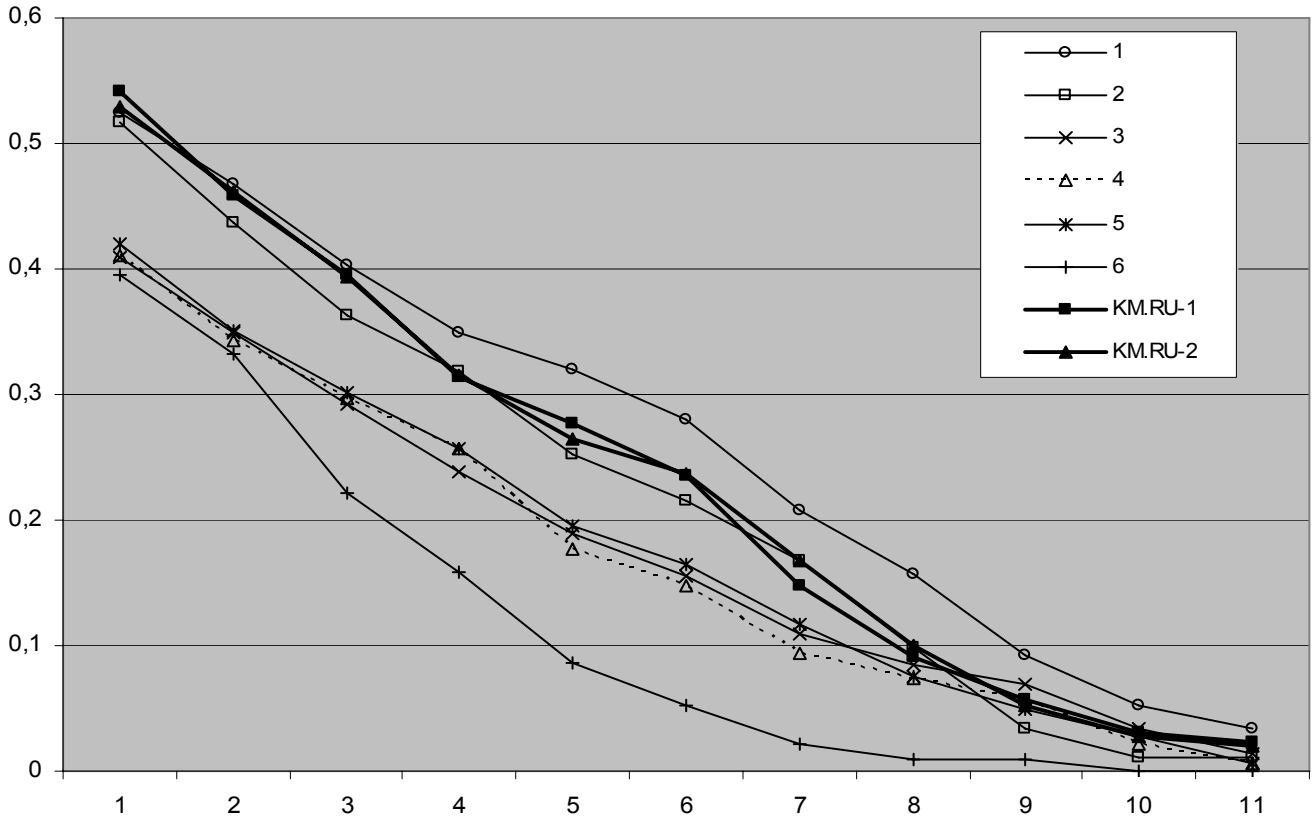
Web adhoc, km.ru, OR



Web adhoc, by.web, AND



Web adhoc, by.web, OR



Выводы

- Высокие по сравнению с другими участниками оценки Precision(5), Precision(10); высокое начало графика TREC: алгоритм позволяет находить и поднимать в начало выдачи высокорелевантные документы
- Модификация алгоритма с *W3* = *F3(paccm)* более эффективна

Спасибо за внимание!