поиск@mail.ru

Mail.Ru на РОМИП’2007

Михаил Костин, Андрей Проскурин

Mail.ru

Поисковые дорожки

- Веб-коллекция (3 прогона)
- Коллекция нормативно-правовых документов (2 прогона)
- Смешанная коллекция (1 прогон)

поиск@mail.ru

Функция релевантности

$$W = K_f W_f + K_p W_p + K_{ps} W_{ps}$$

W_f Частотность термов запроса по TF*IDF

W_p Встречаемость пар соседних слов запроса

W_{ps} Вес наилучшего пассажа в документе

поиск@mail.ru

Пассаж

Запрос: *“Типы характера человека”*

Те сочетания личностных черт, которые входят в [**характер человека**, не являются случайными. Они образуют четко различимые **типы**], позволяющие выявлять и строить…

поиск@mail.ru

Пассажи с «джокерами»

Запрос: *“Значение имени Анастасия”*

<TITLE> [*Значения женских имен*] </TITLE>

<BODY>

…

[*Анастасия*] – *воскресшая (греч.)*

…

</BODY>

поиск@mail.ru

Эксперименты РОМИП

WEB коллекция

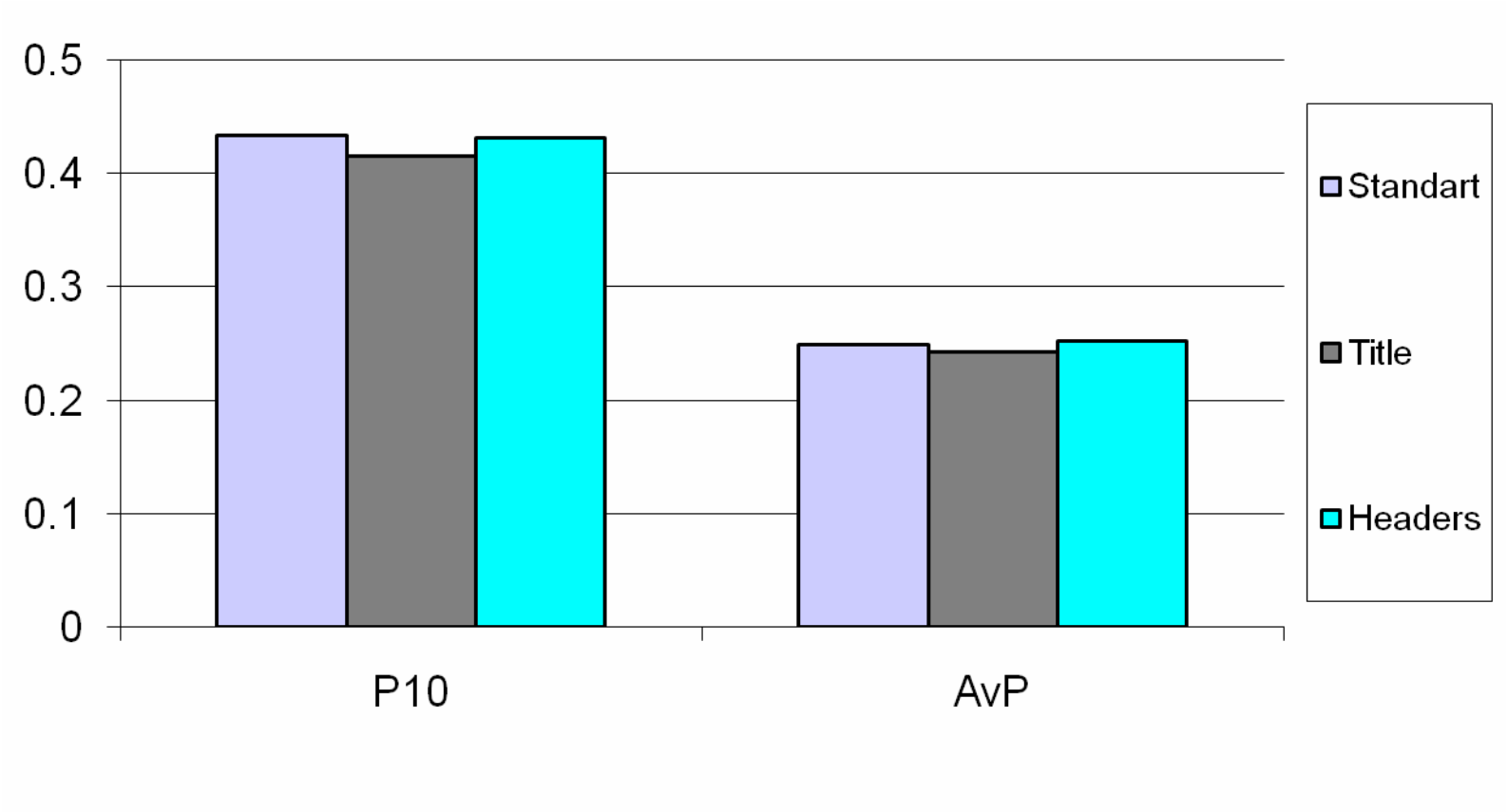
- Обычные пассажи
- Пассажи с «джокерами» из заголовков (title)
- Пассажи с «джокерами» из заголовков, тегов h1-h3 и начала текста документа (первые 100 слов)

LEGAL коллекция

1. Обычные пассажи
2. Пассажи с «джокерами» из заголовков (title)

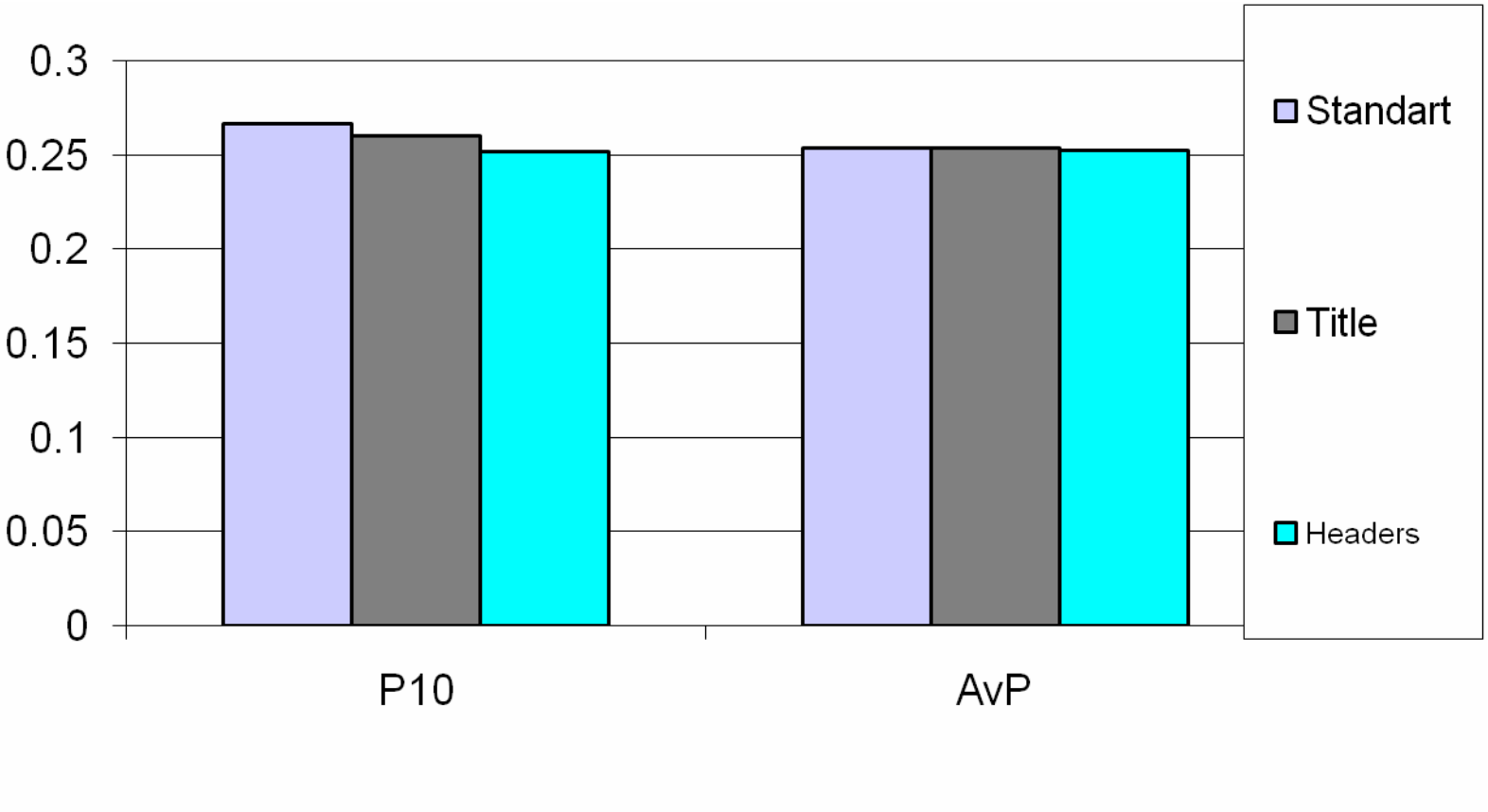
поиск@mail.ru

WEB adhoc ALL – OR, pd50



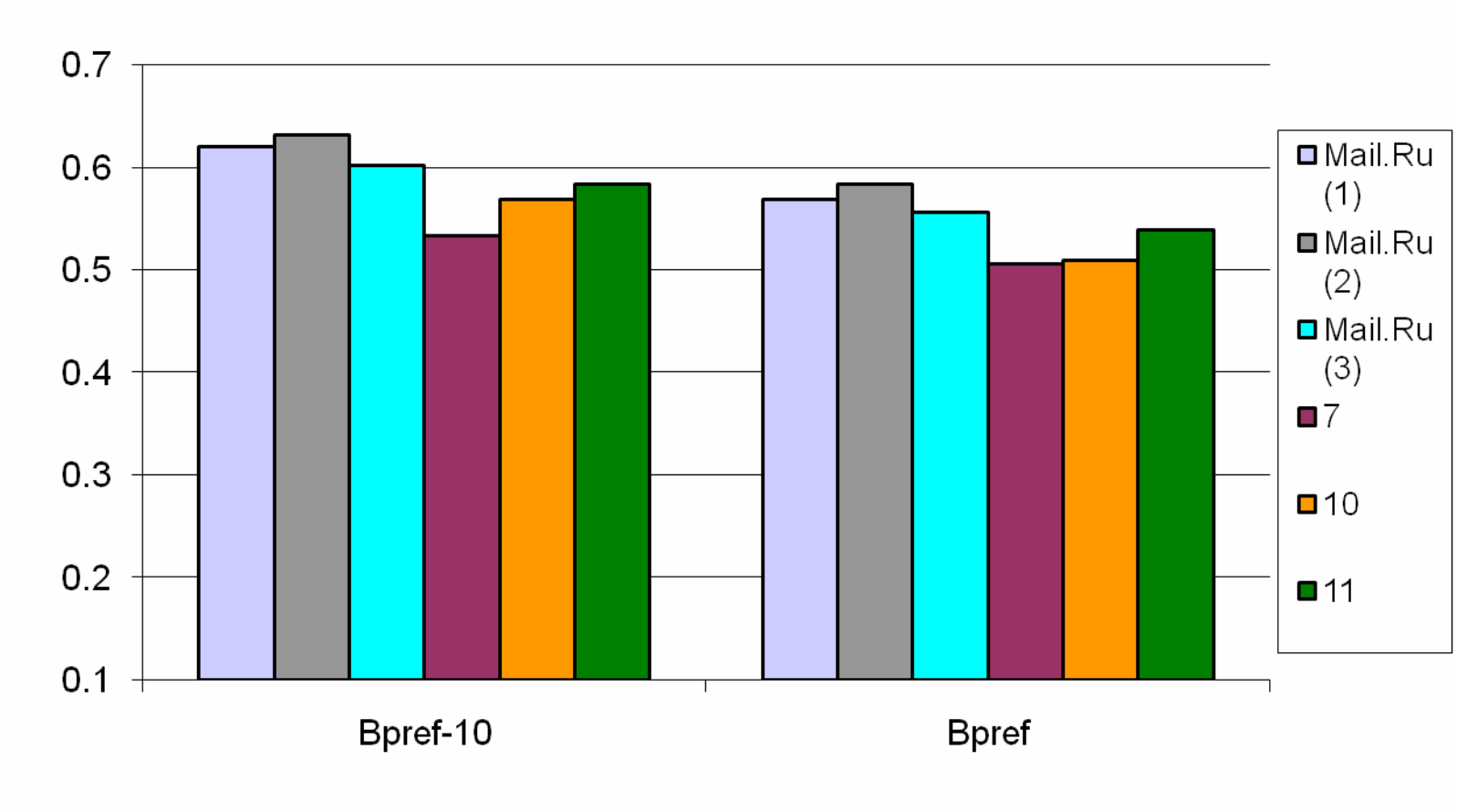
поиск@mail.ru

WEB adhoc ALL - AND, pd50



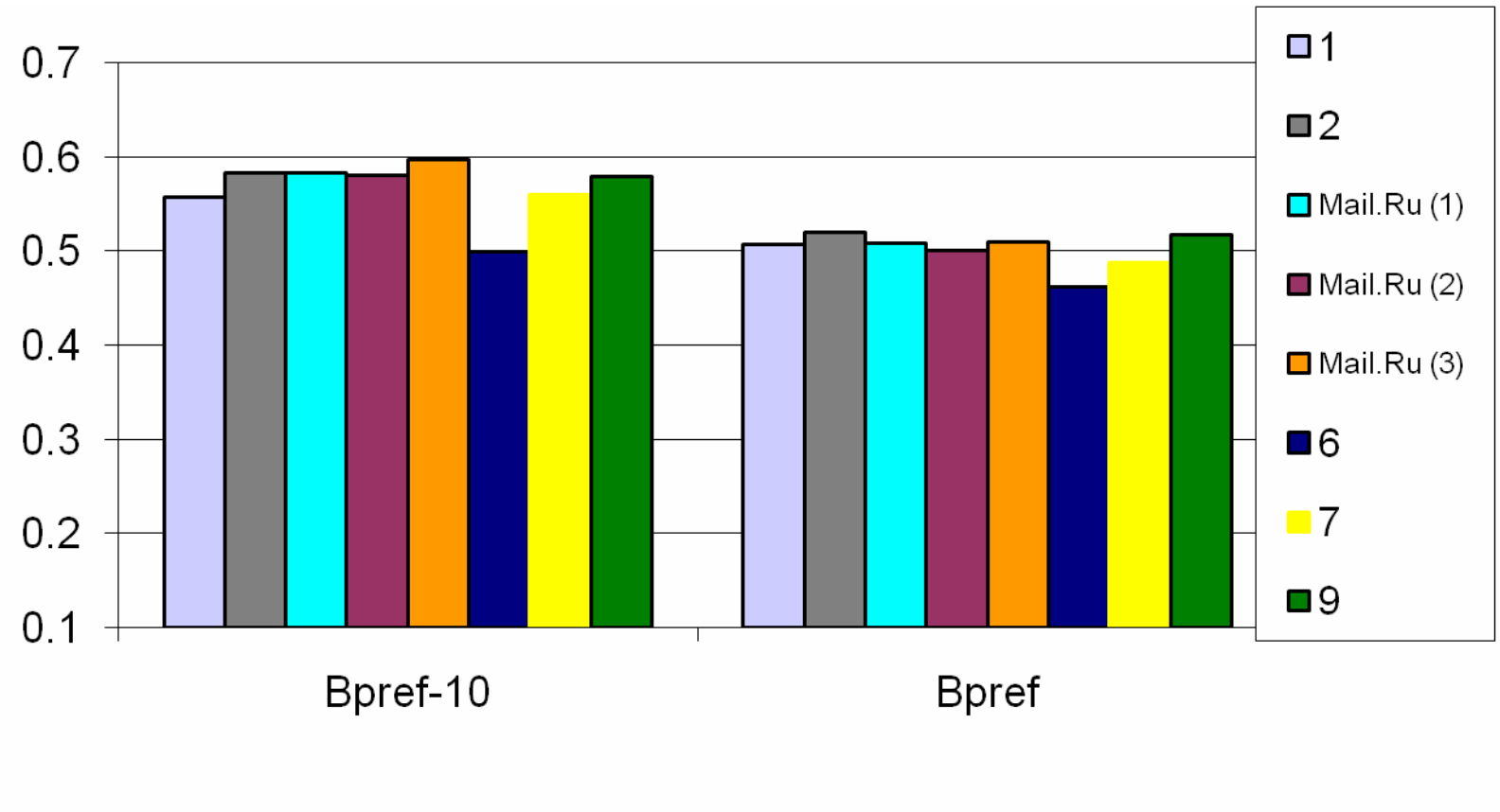
поиск@mail.ru

Web adhoc, KM, AND, pd50



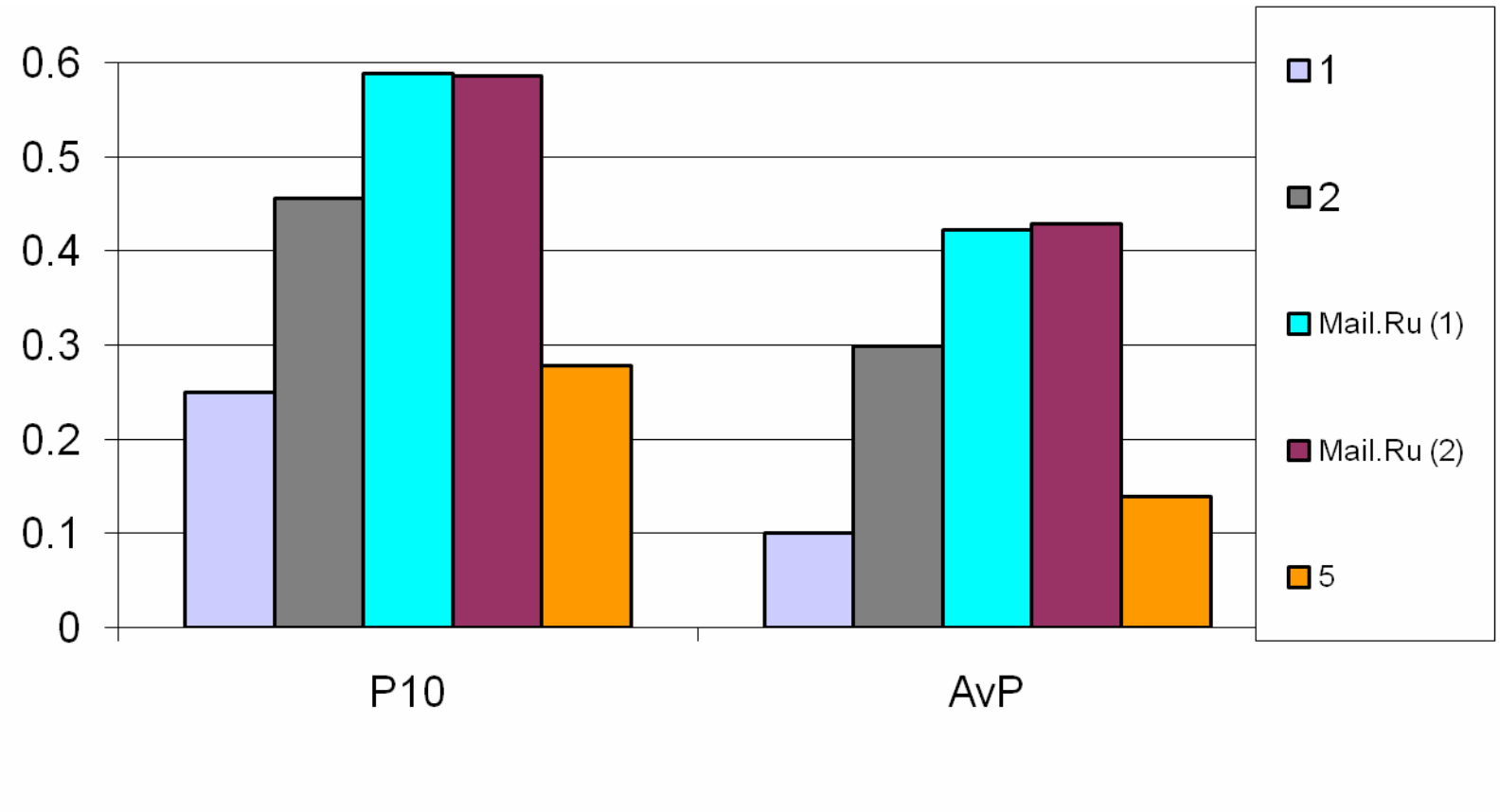
поиск@mail.ru

Web adhoc - BY, AND, pd50



поиск@mail.ru

Legal adhoc - AND, pd50



поиск@mail.ru

Спасибо за внимание !

Mail.Ru,

kostin@corp.mail.ru