# Использование спектральных характеристик лексем для улучшения поисковых алгоритмов

**Зябрев Илья Николаевич**

генеральный директор, AlterTrader Research Ltd.

# Спектральные характеристики лексем

- **Обратная условная частота**

$$ICLF(L,v) = \frac{DF(L)}{CLF(L,v)}$$

- DF(L) – количество документов коллекции, в которых встречается лемма L
- CLF(L,v) – число документов коллекции, в которые лемма L входит v раз.

# Спектральные характеристики лексем

$$SLM(L,v) = \frac{DF(L)}{RCLF(L,v)}$$

- где RCLF(L,v) – число документов коллекции, в которых лемма L имеет относительную частоту равную v.
- **Относительная частота**

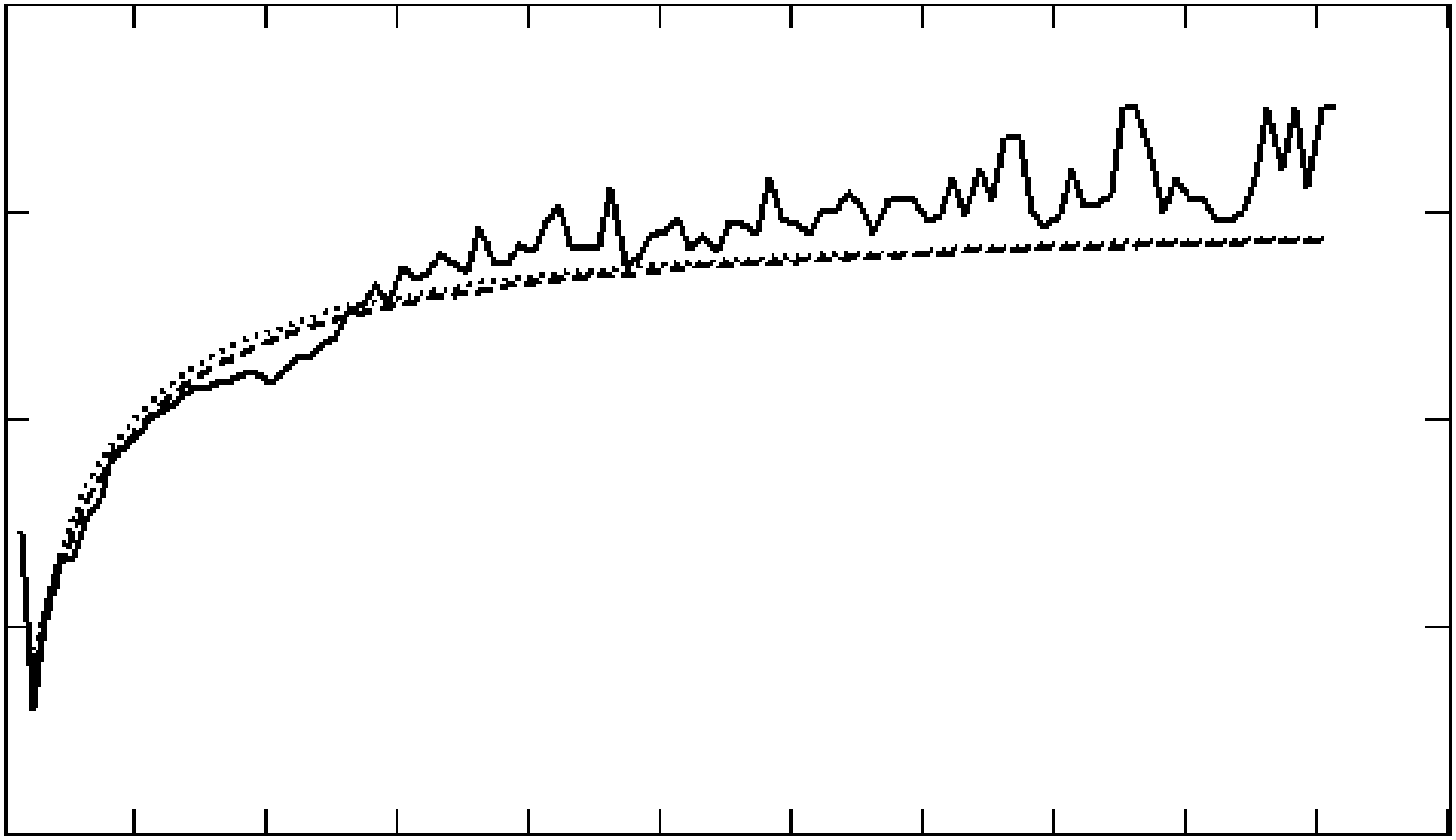
$$RTF(L,d) = \frac{TF(L,d)}{len(d)}$$

- TF(L,d) - внутренняя частота леммы L в документе d,
- len(d) – длина документа d

# Создание базы SLM

- Основа: Коллекции документов KM.RU-2007, BY.WEB-2007;
- Разбиение непрерывной области значений RTF [0, 0.5] на дискретные интервалы;
- Добавочный интервал для значений RTF превышающих 0.5;
- Задача оптимального разбиения не решалась, было выбрано произвольное значение 500 интервалов + 1 добавочный

# Сравнение поведения SLM и BM25



# Сравнение поведения SLM и BM25

- В целом характер поведения у функций является схожим: резкий рост при увеличении относительной частоты на малых значениях и постепенное его замедление  на высоких;
- Однако график log(SLM) имеет ломаный вид, т.к. локально незначительное увеличение частоты может привести к уменьшению значения функции;
- Аналогичная картина на других леммах;

# Базовая ранжирующая формула

За основу взят алгоритм, показавший лучшие результаты на РОМИП-2009

$$Rang(q,d) = k_{doc}M_{doc}(q,d) + k_{title}M_{title}(q,d) + k_{begin}M_{begin}(q,d) + k_{prox}M_{prox}(q,d) + k_{phrase}M_{phrase}(q,d)$$

- kdoc=1, ktitle=2, kbegin=1,5, kprox=1,2, kphrase=10 – коэффициенты, значения которых одинаковы для всех трех реализаций алгоритма.
- q - запрос, d – оцениваемый документ;
- Mdoc(q,d) – вклад всего документа в его ранг;
- Mtitle(q,d) – вклад заголовка документа;
- Mbegin(q,d) – вклад начальной части документа;
- Mprox(q,d) – вклад «кучности» документа;
- Mphrase(q,d) – вклад полноты содержания запроса в документе.

# Алгоритм на базе BM25

$$M(q,d) = BM25(q,d)$$

- **Кучность**

$$M_{prox}(q,d) = \log(1 + \sum_{L \in q} ATC(L,d) \cdot IDF(L))$$

$$ATC(L,d) = \sum_{p \in P(L,d)} \sum_{L' \in q} (\frac{IDF(L)}{LMD(p,L',d)^z} + \frac{IDF(L)}{RMD(p,L',d)^z}) \cdot ts(L,L')$$

- P(L,d) – позиция леммы L в документе d,
- LMD(p, L, d) – расстояние от позиции p до ближайшей слева леммы L в документе d,
- RMD(p, L, d) – расстояние от позиции p до ближайшей справа леммы L в документе d

$$ts(L,L') = \begin{cases} 0.25, L = L' \\ 1, L \neq L' \end{cases}$$

# Алгоритм на базе ICLF

- Получен из базового путем замены IDF на ICLF

$$M(q,d) = \sum_{L \in q} \log(ICLF(L)) \frac{TF(L,d)}{TF(L,d) + 2 \cdot (0.25 + 0.75 \cdot \frac{len(d)}{AvgLen})}$$

- Кучность

$$M_{prox}(q,d) = \log(1 + \sum_{L \in q} ATC(L,d) \cdot ICLF(L))$$

$$ATC(L,d) = \sum_{p \in P(L,d)} \sum_{L' \in q} (\frac{ICLF(L)}{LMD(p,L',d)^z} + \frac{ICLF(L)}{RMD(p,L',d)^z}) \cdot ts(L,L')$$

# Алгоритм на базе SLM

- Получен из базового путем замены BM25 на log(SLM), а IDF на SLM

$$M(q,d) = \sum_{L \in q} \log(SLM(L))$$

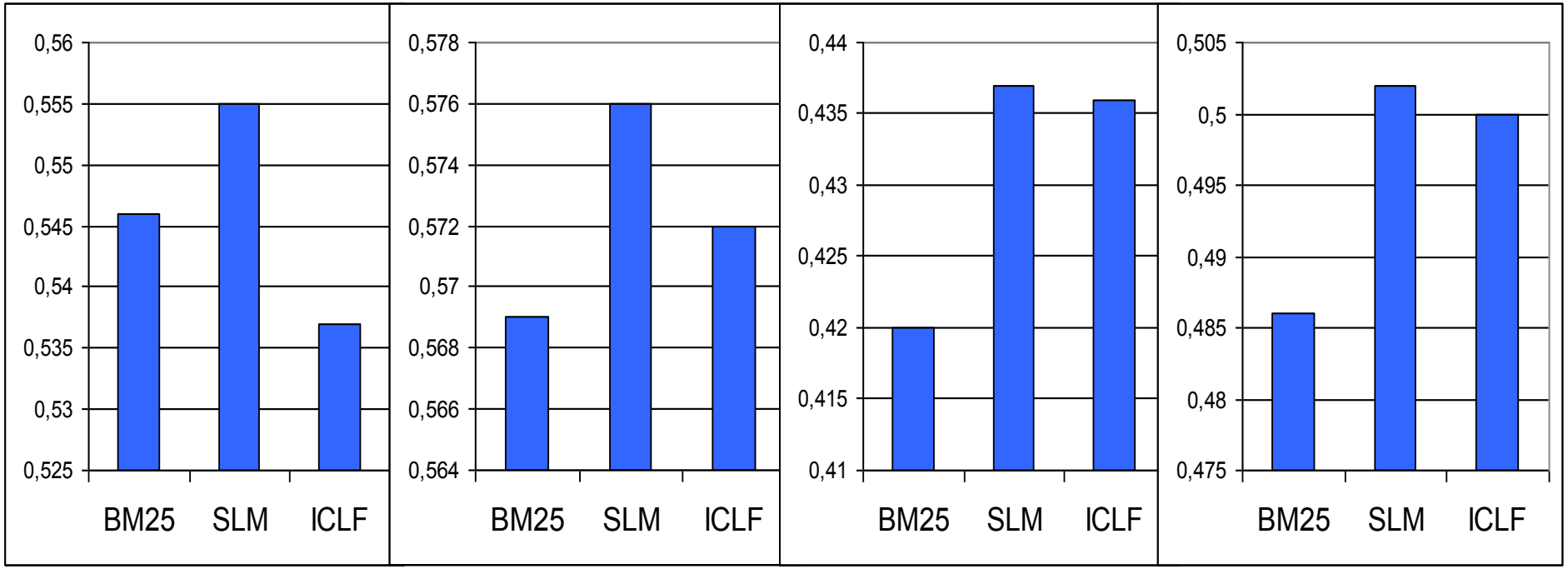
- Кучность

$$M_{prox}(q,d) = \log(1 + \sum_{L \in q} ATC(L,d) \cdot SLM(L))$$

$$ATC(L,d) = \sum_{p \in P(L,d)} \sum_{L' \in q} \left(\frac{SLM(L)}{LMD(p,L',d)^z} + \frac{SLM(L)}{RMD(p,L',d)^z}\right) \cdot ts(L,L')$$

# Сравнение алгоритмов на дорожке поиска по коллекции KM.RU-2007

Метрика pfound по четырем видам оценок:



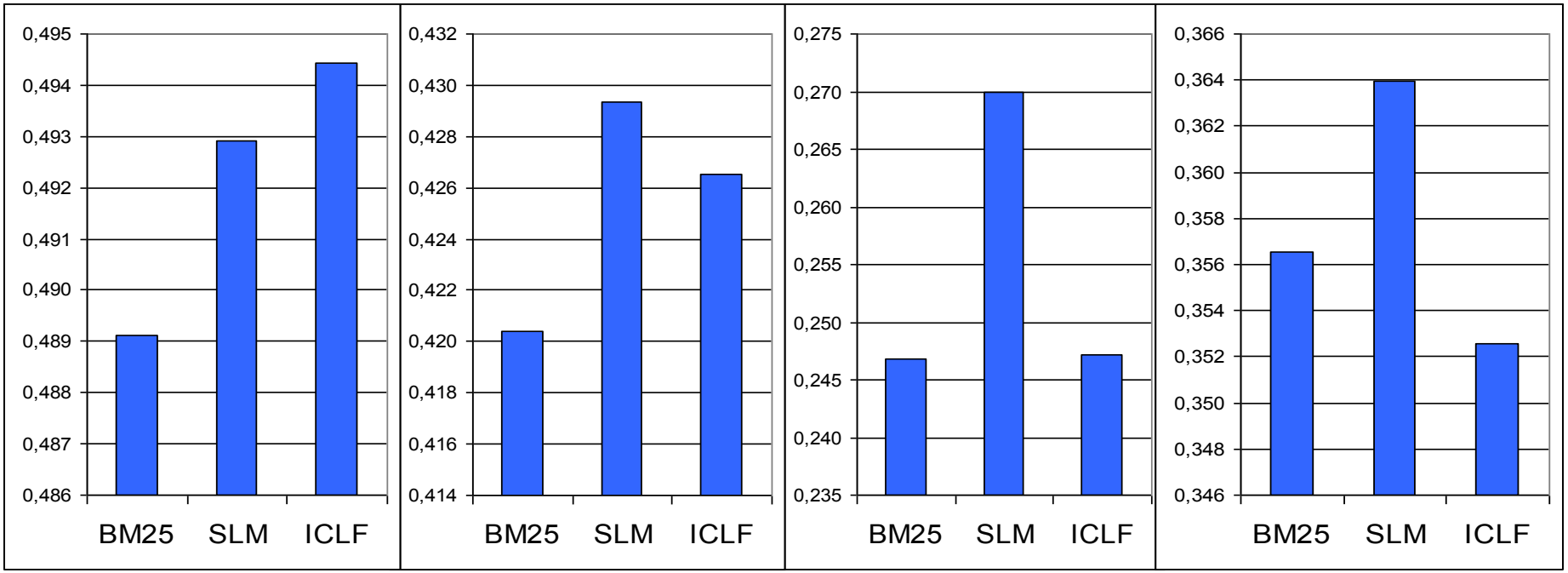

relevance minus and | relevance minus or | relevance plus and | relevance plus or

# Сравнение алгоритмов на дорожке поиска по коллекции KM.RU-2007

- В большинстве случаев: по 61-й оценки из 76 (более 80%) реализация системы на SLM была лучше, чем базовая версия алгоритма на BM25.
- **Наилучшие результаты среди всех представленных систем:**
- **SLM**- версия алгоритма по 24 оценкам;
- **ICLF**-версия алгоритма по 10 оценкам;
- **BM25**-версия алгоритма по 1 оценке;
- Максимальный прирост оценки SLM-версии по сравнению с BM25 18,75% (Precision-1 relevance plus AND)
- Максимальное ухудшение оценки SLM-версии по сравнению с BM25 -2,68% (Precision relevance plus AND)

# Сравнение алгоритмов на дорожке поиска по коллекции BY.WEB-2007

Метрика pfound по четырем видам оценок:




relevance minus and | relevance minus or | relevance plus and | relevance plus or

AlterTrader Research Ltd.
ALTERTRADER

# Сравнение алгоритмов на дорожке поиска по коллекции BY.WEB-2007

- Во всех случаях реализация системы на SLM была не хуже, чем базовая версия алгоритма на BM25: по 82 оценкам из 84  (более 97%) лучше, по двум результат был одинаковый
- ICLF-реализация была лучше BM25-версии по 54 оценкам из 84 (более 65%)
- Максимальный прирост оценки SLM-версии по сравнению с BM25 14,4% (Precision-10 relevance plus and)
- Максимальное ухудшение оценки SLM-версии по сравнению с BM25 0% (Precision-1 relevance minus AND/OR)

# Направления дальнейших исследований

- **Решение задачи оптимального разбиения области значений относительной внутренней частоты слов на дискретные интервалы:**
- Определение оптимального числа равных интервалов.
- Неравномерное разбиение области значений на окрестности фиксированного размера около различных значений RTF слов в документе.
- Неравномерное разбиение области значений на окрестности около различных значений RTF слов в документе. Размер окрестностей имеет функциональную зависимость от одной из частотных характеристик.
- **Построение условных распределений для биграмм и их анализ.**
- **Исследование эффекта использования спектральных характеристик на англоязычных документах**

# Ваши вопросы