Использование спектральных характеристик лексем для улучшения поисковых алгоритмов

Зябрев. И.Н. Пожарков О.В. Пожаркова И.Н.

AltertraderResearch Ltd.,
info@altertrader.com

Аннотация

Статья посвящена использованию спектральных характеристик лексических единиц как основы для поисковых алгоритмов с целью улучшения их качества. Приводятся результаты сравнения поисковой модели на базе BM25 с ее модификациями путем замены классических частотных характеристик на спектральные.

1. Введение

В задачах лингвистического моделирования одной из проблем является оценивание «важности» лексических единиц. На текущий момент наиболее популярной и широко используемой для этих целей метрикой является IDF (Inverse Document Frequency), а также различные функции от нее. Основным недостатком данной оценки является ее независимость от частоты слова внутри документа. Частично данная проблема решается использованием TF*IDF, где TF - относительная частота слова внутри оцениваемого документа, но при этом частота слова в других документах не учитывается. Нами в [1] была предложена спектральная метрика Inverse Conditional Lemm Frequency (ICLF), учитывающая внутренние частоты слов во всех документах коллекции, что предположительно должно было повысить качество оценивания лексем. Однако данное предположение основывалось только на теоретических рассуждениях, поэтому в этом году было принято решение провести сравнительный анализ поисковых алгоритмов, основанных на классических (IDF) и спектральных (ICLF) метриках.

2. Описание алгоритма

2.1 Спектральные характеристики лексем

В [1] были предложены спектральные характеристики лексических единиц, в частности:

Обратная условная частота

$$ICLF(L,v) = \frac{DF(L)}{CLF(L,v)} \quad (1)$$

где DF(L) – количество документов коллекции, в которых встречается лемма L

CLF(L,v) – число документов коллекции, в которые лемма L входит v раз.

В дальнейшем была предложена новая условная характеристика, основанная на внутренней относительной частоте слова:

Спектральная характеристика лексемы (Spectral Lexeme Metrics)

$$SLM(L,v) = \frac{DF(L)}{RCLF(L,v)} \quad (2)$$

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

Относительная частота

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

где TF(L,d) - внутренняя частота леммы L в документе d,

len(d) – длина документа d

На основе коллекций документов KM.ru-2007 и BM.web-2007 для каждой леммы, входящей в их состав, нами были построены базы значений спектральных характеристик ICLF и SLM. Т.к. относительная внутренняя частота, которая является аргументом SLM – непрерывная, то, чтобы ограничить размерность базы, мы разбили диапазон значений от 0 до 0,5 на 500 равных интервалов и один добавочный интервал для случаев, когда значение RTF превышает 0,5, что случается крайне редко. Значение 500 интервалов было выбрано из соображений обеспечения достаточной разрешающей способности по относительной частоте при небольшом объеме базы. Скорее всего, такое разбиение не является оптимальным, однако позволяет решать поисковые задачи.

SLM, по сравнению с обычной ICLF, косвенно учитывает еще и длину документов, что позволяет использовать ее в новом качестве. В частности, было предложено использовать ее как самостоятельную ранжирующую характеристику наряду с BM25.

Рассмотрим особенности поведения SLM в сравнении с BM25 на одной и той же лемме местоимения «Я» (рисунок 1).

На рисунке 1 приведены в едином масштабе графики log(SLM) (сплошная линия), BM25 при фиксированной длине документа равной средней длине по коллекции (пунктирная линия) и BM25 при длине документа равной половине средней (штриховая линия) леммы «Я» при изменении относительной частоты слова от 0 до 0,5.

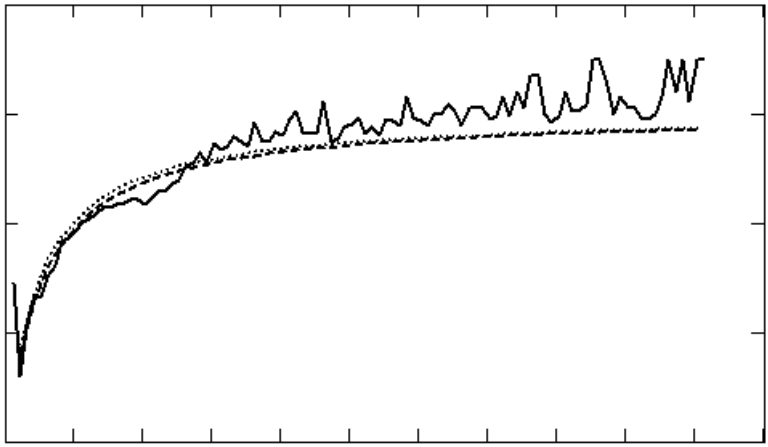

Рисунок 1. Графики log(SLM) и BM25 леммы «Я»

Как видно, 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. На других леммах наблюдается аналогичная картина. Таким образом, log(SLM) имеет схожее с BM25 поведение, однако при этом учитывает особенности распределения внутренних частот лемм по коллекции документов. Поэтому была выдвинута гипотеза о том, что ранжирующие формулы, построенные на основе условных обратных частот, дадут более качественное решение поисковых задач. Для проверки гипотезы было реализовано 3 версии поискового алгоритма.

2.2 Описание ранжирующих алгоритмов

В качестве базового алгоритма была принята небольшая модификация системы, показавшей на прошлой конференции один из лучших результатов [3]. Базовая ранжирующая формула, использованная в каждой из реализаций, имеет вид:

$$Rang(q,d) = k_{doc}M_{doc}(q,d) + k_{title}M_{title}(q,d) + k_{begin}M_{begin}(q,d) + \\ + k_{prox}M_{prox}(q,d) + k_{phrase}M_{phrase}(q,d) \quad (3)$$

Где k_{doc}=1, k_{title}=2, k_{begin}=1,5, k_{prox}=1,2, k_{phrase}=10 – коэффициенты, значения которых одинаковы для всех трех реализаций алгоритма.

q - запрос, d – оцениваемый документ;

$M_{doc}(q,d)$ – вклад всего документа в его ранг;

$M_{title}(q,d)$ – вклад заголовка документа;

$M_{begin}(q,d)$ – вклад начальной части документа;

$M_{prox}(q,d)$ – вклад «кучности» [3] документа;

$M_{phrase}(q,d)$ – вклад полноты содержания запроса в документе.

Каждая из указанных характеристик, за исключением полноты содержания запроса, вычисляется для различных реализаций алгоритма по-разному.

Базовый алгоритм. Первые три характеристики вычисляются по формуле BM25 [4].

$$M(q,d) = \sum_{L \in q} \log(IDF(L)) \frac{TF(L,d)}{TF(L,d) + 2 \cdot (0.25 + 0.75 \cdot \frac{len(d)}{AvgLen})} \quad (4)$$

где IDF(L) – обратная частота леммы L,

AvgLen – средняя длина документа в коллекции.

Кучность оценивается по формуле из [3]

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

$$ATC(L,d) = \sum_{p \in P(L,d)} \sum_{L' \in q} (\frac{IDF(L)}{LMD(p,L',d)^z} + \frac{IDF(L)}{RMD(p,L',d)^z}) \cdot ts(L,L') \quad (5)$$

где 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,

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

Полнота содержания запроса в документе - дискретная величина, которая определяется следующим образом:

$M_{phrase}(q,d)$=4, если запрос содержится полностью в документе, при этом леммы слов запроса идут в документе подряд.

$M_{phrase}(q,d)$=3, если запрос содержится полностью в документе, при этом леммы слов запроса находятся в одном предложении документа.

$M_{phrase}(q,d)=2$, если запрос содержится полностью в документе и указанные выше условия не соблюдаются.

$M_{phrase}(q,d)=1$, во всех остальных случаях.

Модификация алгоритма с использованием ICLF. Все формулы получены из (2)-(3) заменой IDF(L) на ICLF(L):

$$M(q,d)=\sum_{L\in q}\log(ICLF(L))\frac{TF(L,d)}{TF(L,d)+2\cdot(0.25+0.75\cdot\frac{len(d)}{AvgLen})} \quad (6)$$

$$M_{prox}(q,d)=\log(1+\sum_{L\in q}ATC(L,d)\cdot ICLF(L))$$
$$ATC(L,d)=\sum_{p\in P(L,d)}\sum_{L'\in q}(\frac{ICLF(L)}{LMD(p,L',d)^z}+\frac{ICLF(L)}{RMD(p,L',d)^z})\cdot ts(L,L') \quad (7)$$

Модификация алгоритма с использованием SLM.

$$M(q,d)=\sum_{L\in q}\log(SLM(L)) \quad (8)$$

$$M_{prox}(q,d)=\log(1+\sum_{L\in q}ATC(L,d)\cdot SLM(L))$$
$$ATC(L,d)=\sum_{p\in P(L,d)}\sum_{L'\in q}(\frac{SLM(L)}{LMD(p,L',d)^z}+\frac{SLM(L)}{RMD(p,L',d)^z})\cdot ts(L,L') \quad (9)$$

3. Анализ результатов

Проверка алгоритмов проводилась на двух дорожках:

- Поиск по коллекции KM.ru 2007;
- Поиск по коллекции BY.web 2007;

Ниже представлены результаты по коллекции KM.RU (таблицы 1-4)

Таблица 1. Результаты оценки KM.ru relevance minus and

	xxx-1	xxx-2	xxx-3	xxx-4	**BM25**	**SLM**	**ICLF**
Precision(10)	0,510	**0,523**	0,473	0,455	0,516	0,518	0,501
PFound	0,530	0,533	0,491	0,477	0,546	**0,555**	0,537
Graded NDCG@10	0,513	**0,518**	0,455	0,399	0,496	0,504	0,490
NDCG@5	0,227	0,235	0,219	0,196	0,249	**0,254**	0,246
Reciprocal Rank	0,625	0,618	0,569	0,587	0,659	**0,680**	0,667
Graded DCG@10	5,687	5,646	5,228	5,085	5,579	**5,721**	**5,721**
Precision(5)	0,559	0,567	0,523	0,499	0,581	**0,589**	0,564

Precision(1)	0,548	0,534	0,493	0,521	0,575	0,603	**0,616**
Bpref-10	0,528	**0,533**	0,474	0,385	0,512	0,516	0,485
Bpref	0,455	**0,473**	0,403	0,336	0,459	0,459	0,431
DCG@5	1,645	1,662	1,537	1,484	1,708	**1,748**	1,695
Recall	0,777	**0,778**	0,761	0,611	0,724	0,728	0,712
NDCG@10	0,316	0,328	0,309	0,271	0,335	**0,340**	0,327
Graded NDCG@5	0,501	**0,507**	0,433	0,392	0,490	0,503	0,486
Precision	**0,267**	0,264	0,217	0,192	0,252	0,252	0,242
Average precision	0,481	**0,490**	0,431	0,350	0,468	0,473	0,451
DCG@10	2,384	2,431	2,217	2,144	2,435	**2,468**	2,402
Graded DCG@5	4,037	3,950	3,603	3,505	3,859	**4,049**	4,010
R-precision	0,441	0,456	0,409	0,347	**0,458**	0,449	0,424

Таблица 2. Результаты оценки KM.ru relevance minus or

	xxx-1	xxx-2	xxx-3	xxx-4	**BM25**	**SLM**	**ICLF**
Precision(10)	**0,601**	**0,601**	0,529	0,479	0,564	0,569	0,558
PFound	**0,599**	**0,599**	0,534	0,498	0,569	0,576	0,572
Graded NDCG@10	0,513	**0,518**	0,455	0,399	0,496	0,504	0,490
NDCG@5	0,215	**0,220**	0,202	0,159	0,203	0,205	0,211
Reciprocal Rank	0,713	0,723	0,637	0,614	0,709	0,723	**0,731**
Graded DCG@10	5,687	5,646	5,228	5,085	5,579	**5,721**	**5,721**
Precision(5)	**0,643**	**0,643**	0,560	0,510	0,609	0,620	0,616
Precision(1)	0,640	0,652	0,573	0,551	0,663	0,685	**0,697**
Bpref-10	0,554	**0,555**	0,487	0,367	0,487	0,484	0,474
Bpref	0,504	0,512	0,439	0,341	0,459	0,457	0,444
DCG@5	1,901	**1,906**	1,662	1,524	1,818	1,862	1,865
Recall	**0,752**	0,749	0,737	0,574	0,633	0,636	0,630
NDCG@10	0,306	**0,309**	0,283	0,223	0,282	0,285	0,284
Graded NDCG@5	0,501	**0,507**	0,433	0,392	0,490	0,503	0,486
Precision	**0,349**	0,345	0,290	0,246	0,319	0,320	0,313
Average precision	0,510	**0,516**	0,456	0,338	0,452	0,453	0,443
DCG@10	2,798	**2,803**	2,465	2,241	2,652	2,693	2,668

Graded DCG@5	4,037	3,950	3,603	3,505	3,859	**4,049**	4,010
R-precision	0,500	**0,512**	0,451	0,356	0,455	0,449	0,433

Таблица 3. Результаты оценки KM.ru relevance plus and

	xxx-1	xxx-2	xxx-3	xxx-4	**BM25**	**SLM**	**ICLF**
Precision(10)	**0,363**	0,358	0,316	0,326	0,347	0,353	0,342
PFound	0,435	0,427	0,379	0,388	0,420	**0,437**	0,436
Graded NDCG@10	0,513	**0,518**	0,455	0,399	0,496	0,490	0,504
NDCG@5	0,329	0,314	0,278	0,266	0,316	**0,336**	0,323
Reciprocal Rank	0,493	0,470	0,443	0,457	0,503	**0,540**	0,532
Graded DCG@10	5,687	5,646	5,228	5,085	5,579	**5,721**	5,721
Precision(5)	**0,437**	0,428	0,363	0,358	0,372	0,395	0,400
Precision(1)	0,372	0,349	0,349	0,349	0,372	**0,442**	0,419
Bpref-10	**0,539**	0,532	0,481	0,459	0,514	0,522	0,522
Bpref	0,433	0,422	0,383	0,356	0,416	**0,437**	0,409
DCG@5	**1,251**	1,214	1,063	1,046	1,091	1,186	1,179
Recall	**0,847**	**0,847**	0,819	0,750	0,816	0,812	0,814
NDCG@10	0,416	0,401	0,368	0,369	0,415	**0,435**	0,416
Graded NDCG@5	0,501	**0,507**	0,433	0,392	0,490	0,486	0,503
Precision	**0,119**	**0,119**	0,112	0,106	0,104	0,101	0,103
Average precision	**0,485**	0,474	0,436	0,400	0,455	0,466	0,457
DCG@10	**1,721**	1,684	1,502	1,520	1,608	1,689	1,639
Graded DCG@5	4,037	3,950	3,603	3,505	3,859	4,010	**4,049**
R-precision	0,436	0,418	0,381	0,356	0,439	**0,456**	0,438

Таблица 4. Результаты оценки KM.ru relevance plus or

	xxx-1	xxx-2	xxx-3	xxx-4	**BM25**	**SLM**	**ICLF**
Precision(10)	0,427	0,430	0,410	0,394	0,442	0,446	**0,448**
PFound	0,489	0,488	0,445	0,450	0,486	**0,502**	0,500
Graded NDCG@10	0,513	**0,518**	0,455	0,399	0,496	0,504	0,490
NDCG@5	**0,301**	0,298	0,256	0,242	0,282	0,285	0,284
Reciprocal Rank	0,600	0,578	0,507	0,528	0,575	**0,616**	0,615
Graded DCG@10	5,687	5,646	5,228	5,085	5,579	5,721	**5,721**
Precision(5)	0,499	**0,507**	0,454	0,433	0,493	0,504	0,499
Precision(1)	0,537	0,507	0,418	0,433	0,493	**0,537**	**0,537**

Bpref-10	**0,520**	0,512	0,476	0,427	0,486	0,486	0,484
Bpref	**0,466**	0,454	0,400	0,343	0,443	0,438	0,434
DCG@5	1,495	1,491	1,312	1,276	1,445	**1,501**	1,487
Recall	0,759	0,747	**0,780**	0,647	0,730	0,728	0,732
NDCG@10	**0,386**	0,382	0,353	0,330	0,372	0,377	0,383
Graded NDCG@5	0,501	**0,507**	0,433	0,392	0,490	0,503	0,486
Precision	**0,190**	0,186	0,166	0,147	0,180	0,177	0,177
Average precision	**0,485**	0,474	0,427	0,365	0,451	0,451	0,447
DCG@10	2,069	2,060	1,904	1,845	2,076	2,129	**2,131**
Graded DCG@5	4,037	3,950	3,603	3,505	3,859	**4,049**	4,010
R-precision	**0,475**	0,461	0,414	0,361	0,445	0,444	0,437

В таблицах жирным выделены ячейки с лучшими результатами и оценки, по которым алгоритм, построенный на базе SLM, был лучшим среди всех представленных систем. Как видно, по многим оценкам данный алгоритм стал первым, в частности по pfound алгоритм был лучшим в 3 из 4 случаев. В целом тестируемые системы показали следующий результат среди всех участников: по 24 оценкам лучшим была SLM-версия алгоритма, по 10 – ICLF-версия, по 1 – BM25. При этом по большинству оценок (более 80%) реализация системы на SLM была лучше, чем базовая версия алгоритма на BM25.

На коллекции BY.web-2007 по 82 оценкам из 84 (более 97%) алгоритм с использованием SLM был лучше исходной системы и в 2 результат был одинаковым. ICLF-реализация в аналогичном сравнении была лучше по 54 оценкам из 84 (более 65%)

4. Заключение

На основе полученных результатов можно сделать вывод, что гипотеза о более качественном решении поисковых задач при использовании ранжирующих формул, построенных на базе спектральных характеристик, подтверждается. Так как простая замена BM25 на log(SLM) дает увеличение большинства оценок. В частности, согласно окончательным результатам по коллекции KM.ru наблюдается максимальное увеличение оценки на 18,75% (Precision-1 relevance plus AND), максимальное ухудшение оценки - 2,68% (Precision relevance plus AND). По окончательным результатам для коллекции BY.web наблюдается максимальное

увеличение оценки на 14,4% (Precision-10 relevance plus and). Данные результаты позволяют охарактеризовать спектральные характеристики лексем как хорошую замену классическим BM25 и IDF.

В заключении отметим, что представленные результаты были получены на базе условных частот, для которой не решалась задача оптимального разбиения непрерывной области значений на дискретные интервалы. Поэтому одной из возможных перспектив развития спектральных характеристик является поиск подобного разбиения.

Литература

[1] Зябрев И.Н., Пожарков О.В. Спектральное оценивание лексических единиц в задачах лингвистического моделирования http://www.altertrader.com/publications16.html

[2] Зябрев И.Н., Пожарков О.В. Метод контекстно-зависимого аннотирования документов на основе спектральных оценок лексем. Труды ROMIP 2009.Санкт-Петербург: НУ ЦСИ. 2009, с 167-174

[3] Сафронов А.В. HeadHunter на РОМИП-2009. Труды ROMIP 2009.Санкт-Петербург: НУ ЦСИ. 2009, с 63-70

[4] Robertson S.,Walker S., Hancock-Beaulieu M., Gatford M. Okapi at TREC-3. In Proceedings of the Third Text Retrieval Conference. 1994

Spectral characteristics of lexemes using for search algorithms improvement

Zyabrev I.N., Pozharkov O.V., Pozharkova I.N.

Paper is devoted spectral characteristics of lexical units using as a basis for search algorithms for improvement of their quality. Comparison results of search model basis on BM25 with its updatings by replacement classical frequency characteristics by the spectral are presented.