Спектральные характеристики в задачах обработки текстовой информации

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

AltertraderResearch Ltd.,
info@altertrader.com

Аннотация

Данная статья посвящена описанию спектрального подхода в задачах обработки текстовой информации и, в частности, для решения задач информационного поиска. Проведено сравнение спектральной модели (Spectral Language Model – SLM) с популярными вероятностными моделями, такими как BM25 и DFR. Также представлена аппроксимированная спектральная модель, которая позволяет избавиться от главного недостатка SLM – громоздкой частотной базы.

1. Описание спектральной модели SLM

В задачах обработки текстовой информации важнейшей проблемой является «взвешивание» лексических единиц. На текущий момент наиболее популярной и широко используемой для этих целей метрикой является IDF (Inverse Document Frequency) и различные функции от нее. Один из основных недостатков данной оценки – 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. В [3] были предложены характеристики, основанные на распределении частот слова по всей коллекции, которые, в частности, позволили повысить качество решения поисковых задач. Наиболее эффективной с этой точки зрения оказалась характеристика, основанная на нормализованной частоте леммы слова.

Поэтому в дальнейшем данная метрика была взята в качестве базовой для спектральной языковой модели – SLM:

$$SLM(L,d) = \log(\frac{M}{SF(L,nTF(L,d))}), \qquad (1)$$

Где:

Нормализованная частота $nTF(L,d) = \frac{TF(L,d)}{len(d)}$;

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

len(d) – длина документа d;

SF(L,v) – спектральная частота слова, число документов коллекции, в которых слово L имеет нормализованную частоту, равную v.

На основе коллекций документов KM.ru-2007 и BY.web-2007 для лемм всех слов были построены частотные базы, которые в дальнейшем использовались для исследования свойств спектральных характеристик, а также для их сравнения с другими частотными метриками.

Подчеркнем основные свойства спектральной модели, которые были выявлены на основе исследований.

1. Характеристика основана на эмпирических вероятностных распределениях слов по документам коллекции, а не на теоретических, как во многих других вероятностных подходах к взвешиванию слов, например в DFR [1].
2. Вес слова определяется уникальным для каждого слова спектром, в отличие от большинства других характеристик, в которых разные слова при одинаковых значениях TF и DF характеристик равнозначны.
3. Немонотонность изменения значений частотного спектра с ростом нормализованной частоты.

Так как для методов информационного поиска, составляющих одну из важнейших областей обработки текстовой информации, существуют доступные массивы данных (коллекции документов и таблицы релевантностей), позволяющие объективно оценить и сравнить различные технологии, то основные исследования спектральных характеристик были сосредоточены именно в области поисковых технологий.

2. Сравнение SLM с другими вероятностными моделями

В [4] на поисковой дорожке РОМИП-2010 было проведено сравнение двух поисковых методов: алгоритма на основе BM25 [2], показавшего лучшие

Труды 13[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2011, Воронеж, Россия, 2011.

результаты на предыдущем семинаре [5] и его модификации, путем замены BM25 на SLM. В результате, практически по всем оценкам качества, ответы SLM-алгоритма оказались лучше BM25-алгоритма. Т.е. простая замена BM25 на SLM в ранжирующем алгоритме дала прирост качества решения задачи информационного поиска. Однако, в сравнении, проведенном на РОМИП-2010, модели BM25 и SLM использовались лишь в виде отдельных факторов, вычисленных по различным структурным элементам документов. Поэтому, для того чтобы сравнить модели без учета влияния других параметров, было проведено дополнительное исследование моделей на основе таблиц релевантностей РОМИП за 2007–2010 гг.

Для каждой сравниваемой модели (DFR, BM25, SLM) было использовано по 2 ранжирующих алгоритма:

– оценка релевантности документа определяется только по исследуемой модели

$$R1(q,d) = \sum_{L \in q} M_{doc}(q,d) \ , \qquad (2)$$

где q – запрос, d – оцениваемый документ.

– оценка релевантности документа определяется по различным структурным элементам документа

$$R2(q,d) = k_{doc} M_{doc}(q,d) + \\ + k_{title} M_{title}(q,d) + k_{begin} M_{begin}(q,d) \ , \qquad (3)$$

где k_{doc}, k_{title}, k_{begin} – коэффициенты, полученные на основе машинного обучения. Обучение проводилось независимо для каждой модели на основе таблиц релевантностей.

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

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

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

– для SLM: $M(q,d) = \sum_{L \in q} SLM(L,d)$;

– для BM25: $M(q,d) = \sum_{L \in q} BM25(L,d)$;

– для DFR: $M(q,d) = \sum_{L \in q} DFR(L,d)$.

Полученные по каждому алгоритму ответы на запросы оценивались по таблицам релевантностей. Результаты оценок представлены в табл. 1–2.

Таблица 1

Результаты сравнения алгоритмов R1

Evaluation\Systems	DFR	BM25	SLM
Average precision	0,224	0,226	0,256
Bpref	0,551	0,555	0,595
Bpref-10	0,64	0,643	0,685
Precision(1)	0,454	0,472	0,522
Precision(10)	0,442	0,46	0,51
Precision(5)	0,444	0,464	0,514
Reciprocal Rank	0,458	0,48	0,53
R-precision	0,28	0,296	0,32
NDCG@5	0,242	0,257	0,282
DCG@5	0,835	0,863	0,961
NDCG@10	0,330	0,339	0,366
DCG@10	1,306	1,315	1,451

Таблица 2

Результаты сравнения алгоритмов R2

Evaluation\Systems	DFR	BM25	SLM
Average precision	0,26	0,266	0,296
Bpref	0,678	0,685	0,748
Bpref-10	0,782	0,788	0,858
Precision(1)	0,522	0,538	0,588
Precision(10)	0,512	0,53	0,576
Precision(5)	0,514	0,53	0,58
Reciprocal Rank	0,322	0,34	0,357
R-precision	0,526	0,542	0,597
NDCG@5	0,379	0,387	0,435
DCG@5	1,203	1,231	1,406
NDCG@10	0,467	0,478	0,524
DCG@10	1,772	1,802	2,026

Как видно из таблиц, по обоим алгоритмам лучшие результаты по всем оценкам получила спектральная модель. В среднем оценки SLM выше на 10% по сравнению с BM25 и на 13% по сравнению с DFR, что считается существенной разницей. На рис. 1 представлен график TREC для ответов по алгоритму R1.

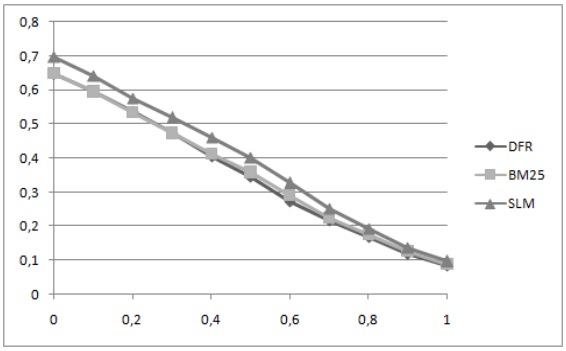

Рисунок 1. График TREC ответов по алгоритму R1.

Из графика также видно, что точность результатов поиска при одинаковых значениях полноты у алгоритма на основе SLM выше, по сравнению с DFR и BM25.

3. Аппроксимированная SLM

В целом полученные результаты позволяют говорить о том, что спектральная модель, по крайней мере на русскоязычных документах, дает более качественное решение поисковых задач по сравнению с другими методами. Однако спектральная модель обладает существенным недостатком – очень большой размер частотной базы. Если в большинстве вероятностных моделей на каждое слово в частотную базу заносится не более двух параметров, то здесь их число существенно больше. Один из способов уменьшения базы - выбор большего шага частотной дискретизации. Однако данный метод не решает полностью проблему размера частотной базы, т.к. уже при шаге больше 0,01, что соответствует разбиению области значений на 100 интервалов, наблюдается снижение качества решения задач на основе SLM.

Проведенные исследования показали, что спектры слов можно аппроксимировать с минимальными потерями качества решения поисковых задач функцией от 3 аргументов aSF(nTF,a,b) где a и b – параметры, которые определяются для каждого слова на основе метода наименьших квадратов. При этом сохраняется свойство уникальности спектра слов, а размер частотной базы существенно сокращается: на каждое слово необходимо хранить по 2 параметра.

Лучший результат из исследованных нами функций показала степенная:

$$aSF(nTF, a, b) = a \cdot nTF^{b}. \qquad (4)$$

Соответствующая ей аппроксимированная SLM (aSLM) с переходом к другим константам имеет вид:

$$aSLM(nTF, a, b) = a + b \cdot \log(nTF). \qquad (5)$$

На основе метода наименьших квадратов для каждого слова были получены и занесены в базу значения параметров. На рисунке 2 изображены графики базовой SLM и аппроксимированной для местоимения «Я».

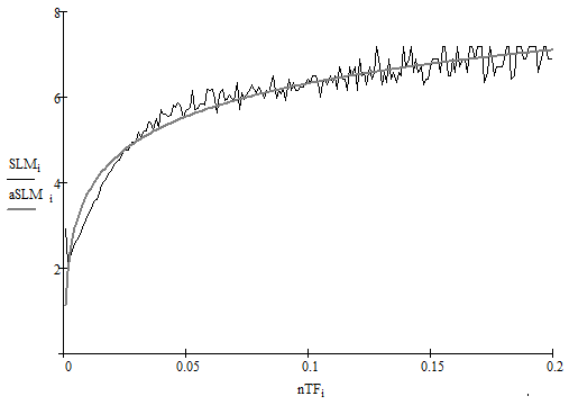

Рисунок. 2. Графики базовой SLM и аппроксимированной SLM местоимения «Я».

Как видно, визуально приближение исходной спектральной модели функцией (5) довольно хорошее. Для того, чтобы оценить, насколько использование аппроксимации ухудшает качество решения поисковых задач, было проведено исследование, аналогичное сравнительному анализу SLM с другими вероятностными моделями. Результаты оценок представлены в таблицах 3–4.

Таблица 3

Результаты сравнения алгоритмов R1

Evaluation\Systems	SLM	aSLM
Average precision	0,256	0,258
Bpref	0,595	0,606
Bpref-10	0,685	0,715
Precision(1)	0,522	0,539
Precision(10)	0,51	0,522
Precision(5)	0,514	0,526
Reciprocal Rank	0,53	0,535
R-precision	0,32	0,321
NDCG@5	0,282	0,284
DCG@5	0,961	1,003
NDCG@10	0,366	0,367
DCG@10	1,451	1,514

Таблица 4

Результаты сравнения алгоритмов R2

Evaluation\Systems	SLM	aSLM
Average precision	0,296	0,311
Bpref	0,748	0,779
Bpref-10	0,858	0,893
Precision(1)	0,588	0,619
Precision(10)	0,576	0,602
Precision(5)	0,58	0,608
Reciprocal Rank	0,357	0,371
R-precision	0,597	0,626
NDCG@5	0,435	0,448
DCG@5	1,406	1,451
NDCG@10	0,524	0,545
DCG@10	2,026	2,087

Из таблиц видно, что aSLM по обоим алгоритмам улучшает качество решения поисковых задач: по алгоритму R1 в среднем на 1%, по алгоритму R2 в среднем на 5%. Таким образом, использование аппроксимированной модели на основе функции (5) не только не ухудшает качество решения поисковой задачи, но и незначительно его улучшает. При этом объем частотной базы сокращается на два порядка.

4. Заключение

Проведенные исследования показали, что спектральная языковая модель позволяет более качественно решать поисковые задачи по сравнению с обычными вероятностными моделями, которые не учитывают особенности распределения различных слов по документам коллекции. Единственным существенным недостатком SLM относительно большинства параметрических моделей является огромный размер частотной базы. Однако, использование аппроксимирующих функций для спектров слов позволяет свести модель к двухпараметрической, уменьшая число хранимых параметров для каждой леммы до 2. Сравнительный анализ aSLM и исходной SLM показал, что качество решения поисковых задач при использовании функции (5) улучшается.

Таким образом, спектральные характеристики являются хорошей альтернативой различным частотным метрикам, используемым в задачах обработки текстовой информации и, в частности, их применение в поисковых алгоритмах позволяет увеличить качество поиска по сравнению с широко распространенными вероятностными методами (BM25, DFR).

Литература

[1] Amati, G. Probabilistic models of information retrieval based on measuring the divergence from randomness / G. Amati and C. J. Van Rijsbergen, The Information Retrieval Group, 20(4):357-389, 2002.

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

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

[4] Зябрев И. Н., Пожарков О. В., Пожаркова И. Н. Использование спектральных характеристик лексем для улучшения поисковых алгоритмов. Труды РОМИП 2010. Казань: Казан. ун-т: С. 40–48, 2010.

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

Spectral Characteristics in Problems of Text Information Processing

© Ilya Zyabrev, Oleg Pozharkov, Irina Pozharkova

Paper is devoted to the description of the spectral approach in problems of the text information processing and, in particular, for the decision of information retrieval problems. Comparison of spectral model (SLM) with popular probability models, such as BM25 and DFR, is provided. Also the approximate spectral model is presented.