S. G. Zablotskiy, E. S. Semenkin, A. V. Shvets

COEVOLUTIONARY ASYMPTOTIC GENETIC ALGORITHM FOR SYLLABLE MODEL BASED IDENTIFICATION OF SENTENCES IN AUTOMATIC RECOGNITION OF CONTINUOUS SPEECH

The last stage of the syllable based automatic continuous speech recognition is the concatenation of the recognized syllables into a sequence of words. A method for solving this problem is suggested and investigated in the article. It is based on a specially designed stochastic optimization algorithm which is able to determine the most likely sentence corresponding to acoustic signal within an acceptable time frame.

Keywords: coevolutionar algorithm, syllable based model, speech recognition, dynamic programming.

УДК 004.738.52

И. Н. Зябрев, К. Н. Василенко, О. В. Пожарков, И. Н. Пожаркова

СРАВНИТЕЛЬНЫЙ АНАЛИЗ СПЕКТРАЛЬНОЙ ЯЗЫКОВОЙ МОДЕЛИ И ДРУГИХ МОДЕЛЕЙ ИНФОРМАЦИОННОГО ПОИСКА

Проведено сравнение спектральной языковой модели (SLM) с наиболее широко используемыми в информационном поиске моделями, такими как DFR (Difference From Randomness) и BM25 с точки зрения оценки качества решения поисковых задач.

Ключевые слова: информационный поиск, вероятностная модель, SLM.

Классическая задача информационного поиска формулируется в следующем виде. Существует множество (коллекция) документов D (d_1, d_2, … d_M); M – его мощность и формализованный запрос q, выражающий информационные потребности пользователя поисковой системы. Требуется найти документы коллекции D, удовлетворяющие запросу q. Решение задачи обычно представляет собой список документов D_{rel}, в порядке убывания их степени соответствия запросу (релевантности).

В настоящее время большинство алгоритмов оценки релевантности документов построено на основе таких моделей, как BM25 [1] и DFR [2]. К основным достоинствам данных моделей можно отнести довольно высокое качество решения поисковых задач, малую вычислительную сложность и небольшой размер частотной базы, необходимой для вычисления оценок релевантности документов. Среди наиболее существенных недостатков большинства моделей информационного поиска, в частности, BM25 и DFR, стоит выделить унифицированное взвешивание отдельных слов документа, которое приводит к тому, что слова, имеющие одинаковые частоты в оцениваемом документе и во всей коллекции, будут давать одинаковый вклад в оценку релевантности, несмотря на их различную информационную значимость. Спектральная языковая модель (SLM) [3] учитывает распределение слов по всем документам коллекции и этим существенно отличается от указанных моделей, позволяя учитывать статистическую значимость слов при оценке релевантности (взвешивании) документа. Для того чтобы оценить, как такое свойство модели влияет на качество решения поисковых задач, был проведен сравнительный анализ SLM и наиболее широко используемых в данный момент вероятностных моделей: BM25 и DFR.

Описание спектральной языковой модели (SLM). Имеется множество слов $W(w_1, w_2, ..., w_N)$ и множество документов коллекции $D(d_1, d_2, … d_M)$, N и M – соответственно их мощности. Для каждой пары «слово–документ» определена нормализованная частота:

$$nTF(w_i, d_j) = TF(w_i, d_j)/\text{len}(d_j),$$

где $TF(w_i,d_j)$ – частота слова w_i в документе d_j; $\text{len}(d_j)$ – длина документа d_j в словах; $i = 1...N$, $j = 1...M$.

Тогда спектральная частота ($SF(w, F)$) слова w представляет собой число документов, в которых w имеет нормализованную частоту, равную F:

$$– SF(w, TF) = \{D:nTF(w, d_j) = F\},$$

где {*} – оператор мощности множества.

Множество спектральных частот, определенных над областью значений $nTF \in [0,1]$, образует частотный спектр слова.

Базовой характеристикой, на основе которой решаются задачи информационного поиска в различных моделях, является вес слова w в документе d. В спектральной модели она определяется по формуле

$$\text{SLM}(w, d) = \log(M/SF(w, nTF(w, d)).$$

Оценка релевантности документа d относительно запроса q определяется по формуле

$$\mathrm{Re}\,l(q,d) = \sum_{L \in q} \mathrm{SLM}(L,d),$$

где L – лемма (каноническая форма) слова запроса.

Так как nTF – величина непрерывная, то для построения базы нормализованных частот область значений была разбита на интервалы длиной 0,001. Такая избыточная разрешающая способность была выбрана с целью решения в дальнейшем задачи оптимального разбиения на интервалы.

Последующие исследования показали, что уменьшение интервала разбиения или усреднение значений спектральных частот из окрестности nTF заданной ширины приводит к ухудшению результатов поиска, однако даже при увеличении шага дискретизации до 0,01 качество решения поисковых задач остается достаточно высоким по сравнению с другими моделями.

Сравнение SLM с вероятностными моделями BM25 и DFR. Для независимого оценивания спектральной модели в рамках российского семинара по оценке методов информационного поиска было проведено сравнение двух поисковых алгоритмов: на базе SLM и BM25 [3]. По каждому из них были получены ответы на более чем 40 тысяч предложенных запросов по коллекции из 5 миллионов документов. Оценка ответов проводилась асессорами по методу общего котла. Алгоритм на основе SLM показал практически по всем видам оценок существенно лучшие результаты по сравнению с BM25 (табл. 1).

Однако в сравнении, проведенном на РОМИП-2010, участвовали модели BM25 и SLM не в чистом виде, так как для улучшения качества ранжирования использовался аддитивный алгоритм оценки релевантности документа с использованием весов, вычисленных по различным структурным элементам документов. Поэтому чтобы исключить влияние подобных факторов, было проведено дополнительное исследование моделей на основе таблиц релевантностей РОМИП за 2007–2010 гг.

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

– оценка релевантности документа определяется только по исследуемой модели:

$$R_1(q,d) = M_{\mathrm{doc}}(q,d),$$

где q – запрос; d – оцениваемый документ;

– оценка релевантности документа определяется по различным структурным элементам документа:

$$R_2(q,d) = k_{\mathrm{doc}} M_{\mathrm{doc}}(q,d) + \\ + k_{\mathrm{title}} M_{\mathrm{title}}(q,d) + k_{\mathrm{begin}} M_{\mathrm{begin}}(q,d),$$

где k_{doc}, k_{title}, k_{begin} – коэффициенты, полученные на основе машинного обучения; обучение проводилось независимо для каждой модели на основе таблиц релевантностей; $M_{\mathrm{doc}}(q, d)$ – вклад всего документа в оценку его релевантности; $M_{\mathrm{title}}(q, d)$ – вклад заголовка документа; $M_{\mathrm{begin}}(q, d)$ – вклад начальной части документа; для SLM: $M(q,d) = \sum_{L \in q} \mathrm{SLM}(L,d)$; для BM25: $M(q,d) = \sum_{L \in q} \mathrm{BM25}(L,d)$; для DFR: $M(q,d) = \sum_{L \in q} \mathrm{DFR}(L,d)$;

Полученные на запросы ответы по каждому алгоритму оценивались по таблицам релевантостей. Результаты оценок представлены в табл. 2–3.

Как видно, по обоим алгоритмам лучшие результаты по всем оценкам получила спектральная модель. В среднем оценки SLM выше на 10 %, чем у BM25, и на 13 %, чем у DFR, что считается существенной разницей. На рисунке представлен график TREC для ответов по алгоритму R_1.

Таблица 1

Результаты сравнения ранжирующих алгоритмов на РОМИП-201

Evaluation\System	BM25	SLM
Average precision	0,455	0,466
Bpref	0,416	0,437
Bpref-10	0,514	0,522
Precision(1)	0,372	0,442
Precision(10)	0,347	0,353
Precision(5)	0,372	0,395
Reciprocal Rank	0,503	0,54
R-precision	0,439	0,456
NDCG@5	0,316	0,336
DCG@5	1,091	1,186
NDCG@10	0,415	0,435
DCG@10	1,608	1,689

Таблица 2

Результаты сравнения алгоритмов R_1

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 3

Результаты сравнения алгоритмов R_2

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

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

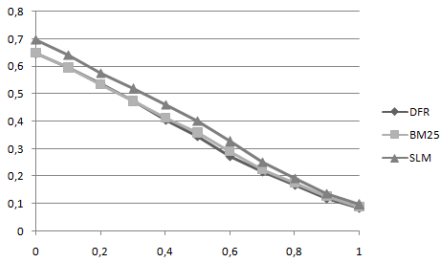

График TREC ответов по алгоритму R_1

В целом, полученные результаты свидетельствуют, что спектральная модель по крайней мере на русскоязычных документах дает более качественное решение поисковых задач.

Проведенные исследования говорят о том, что спектральный подход в оценивании релевантностей документов дает более качественное решение поисковых задач в сравнении с другими методами. Кроме того, SLM является непараметрической моделью, т. е. не требует настройки или машинного обучения, в отличие от многих других моделей, используемых в задачах информационного поиска. Однако спектральная модель обладает существенным недостатком – очень большим размером частотной базы. Если в большинстве вероятностных моделей на каждое слово в частотную базу заносится не более двух параметров, то здесь их число существенно больше. Одним из способов уменьшения базы может быть выбор большего шага дискретизации. Однако проведенные исследования показали, что спектры слов можно аппроксимировать с минимальными потерями качества решения поисковых задач функцией от трех аргументов, в частности, функцией вида $aSF(nTF, a, b) = a \cdot nTF^b$, $b < 0$, 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 два параметра.

Библиографические ссылки

1. Okapi at TREC-3 / S. Robertson, S. Walker, M. Hancock-Beaulieu, M. Gatford // Proc. of the Third Text Retrieval Conf. 1994. P. 109–126.

2. Amati G., Van Rijsbergen C. J. Probabilistic models of information retrieval based on measuring the divergence from randomness // The Information Retrieval Group. 2002. № 20(4). P. 357–389.

3. Зябрев И. Н., Пожарков О. В., Пожаркова И. Н. Использование спектральных характеристик лексем для улучшения поисковых алгоритмов // Тр. РОМИП. 2010. Казань : Изд-во Казан. ун-та, 2010. С. 40–48.

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

THE COMPARATIVE ANALYSIS OF SPECTRAL LANGUAGE MODEL AND OTHER INFORMATION RETRIEVAL MODELS

Comparison of spectral language model (SLM) with models most widely used in information retrieval, such as DFR (Difference From Randomness) and BM25 by quality estimation of the search problems decision.

Keywords: information retrieval, probability model, SLM.

УДК 658.512.001.56

М. В. Карасева, Д. В. Кустов

ИНФОРМАЦИОННО-ТЕРМИНОЛОГИЧЕСКИЙ БАЗИС В МУЛЬТИЛИНГВИСТИЧЕСКОЙ АДАПТИВНО-ОБУЧАЮЩЕЙ ТЕХНОЛОГИИ*

Рассмотрено функциональное назначение и структура программной системы TuMLas v.1,0, реализующей программно-алгоритмическое обеспечение мультилингвистической адаптивно-обучающей технологии. Представлена реляционная модель структуры информационно-терминологического базиса, а также структура первичного информационно-терминологического базиса в виде концептуальной ER-диаграммы.

Ключевые слова: мультилингвистическая технология, программная система, ER-диаграмма.

Одним из ключевых моментов в обучении иностранному языку является изучение иностранной лексики. В этом случае направленная методика обучения подразумевает создание профессионально ориентированных словарей. Если словарь исполнен в классическом виде, то термины в нем располагаются в алфавитной последовательности. Это упрощает их поиск, но никоим образом не повышает эффективность обучения при использовании такого словаря.

С этой целью разработана мультилингвистическая адаптивно-обучающая технология [1], которая рассматривает свойства лексем применительно к конкретным предметным областям, что позволяет для каждой области сформировать информационный терминологический базис (ИТБ) [2]. На основе данных о его элементах становится возможным принять решение о той или иной структуре профессионально ориентированного словаря или же самого ИТБ.

Вопрос об эффективной организации ИТБ нашел решение в ряде методов оптимизации [3], но ни один из этих методов до сих пор не учитывал ассоциативные связи между лексемами (понятиями) того или иного языка.

Такие методы были разработаны и реализованы в программной системе TuMLas v.1,0 [4].

Функциональное назначение программной системы заключается в осуществлении программно-алгоритмической поддержки мультилингвистической адаптивно-обучающей технологии. TuMLas v.1,0 является высококачественным инструментом построения мультилингвистического ИТБ на основе новейших методов оптимизации его структуры [5].

Система работает в двух взаимосвязанных режимах, каждый из которых соответствует ее отдельной функции и обеспечивается отдельным функциональным блоком. Это режим анализа текстов и последующей генерации первичного информационно-терминологического базиса; режим формирования целевого (пригодного для использования в процессе обучения) информационно-терминологического базиса путем реорганизации структуры первичного базиса, в том числе с помощью методов оптимизации его структуры.

Структура разработанной программной системы представлена на рис. 1.

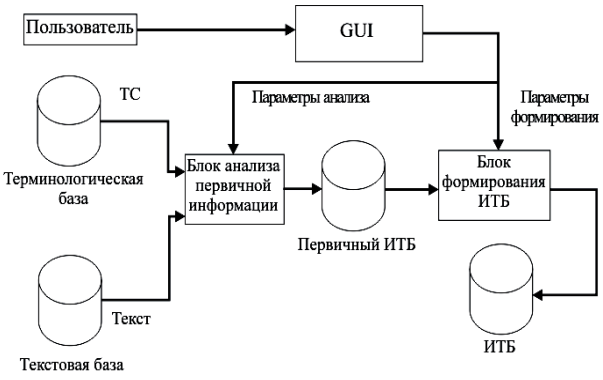

Рис. 1. Структура программной системы TuMLas v.1,0

* Работа выполнена при финансовой поддержке ФЦП «Научные и научно-педагогические кадры инновационной России на 2009–2013 гг.» и ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007–2013 годы» 2011-1.9-519-005.