УДК 681.32

М. В. Карасева

ИСПОЛЬЗОВАНИЯ МЕТОДОЛОГИИ PLSA ДЛЯ АДАПТИВНОЙ КОРРЕКТИРОВКИ МОДЕЛИ ПОЛЬЗОВАТЕЛЯ*

Рассматривается алгоритм непрерывной корректировки модели (профиля) пользователя. Исходными данными являются начальный профиль и история предыдущих запросов. В алгоритме используется методология PLSA. Для достижения поставленной цели вводится понятие временного латентного семантического пространства.

Ключевые слова: вероятностный латентно-семантический анализ, модель пользователя, профиль пользователя.

Потребность в компьютерных системах, осуществляющих автоматический поиск и фильтрацию информации в огромных по объему хранилищах документов (например в Интернете), привела к появлению целой области исследований, связанных с созданием поисковых интерфейсов пользователя [1].

В настоящее время основными инструментами поиска информации в сети являются поисковые сервисы. Поиск информации, как правило, начинается с введения запроса в одном из поисковых сервисов, в ответ на который поисковый сервис в большинстве случаев выдает тысячи документов, после чего полученная информация делится на релевантную (значимую для пользователя) и нерелевантную. Не нарушая общности, здесь и далее будем считать, что информация представляется пользователю в виде текстовых документов.

Индивидуализация (персонализация) интерфейса пользователя благодаря алгоритмам идентификации позволяет учитывать его неявные интересы и использовать их в контексте текущего запроса. Тем самым еще на стадии обработки результатов запроса бóльшая часть нерелевантных документов отсеивается.

Один из распространенных подходов к представлению документов (и запросов) при извлечении информации из сети основан на понятии модели векторного гиперпространства [2], которое в методологии латентной семантической индексации заменяется представлением документа в латентном пространстве меньшей размерности [3]. В данной стате предлагается расширить понятие латентного семантического пространства с учетом текущих интересов пользователя.

Методология PLSA в области извлечения информации. Проблема поиска (извлечения) текстовой информации из ее обширных хранилищ (репозиториев) приобрела особую актуальность в связи с появлением всемирной сети Интернет. В настоящее время каждый пользователь, имеющий доступ в Интернет, может обратиться ко всем источникам информации, представленным в нем. Казалось бы, что теперь своевременное получение необходимой информации по интересующей тематике обеспечиваться без особых затруднений. Однако на деле оказывается, что качество поиска информации при всей ее доступности очень низкое. В поисковых сервисах отсутствуют эффективные алгоритмы поиска релевантной информации, т. е. набора релевантных документов, отражающих суть запроса. И в ответ на запрос такой сервис может выдать сколь угодно большое количество документов, либо отдаленно отражающих сферу интересов пользователя, либо вовсе не имеющих никакой связи с запросом.

Для разрешения проблемы поиска информации могут использоваться два подхода: с одной стороны – традиционный лингвистический подход, сторонники которого пытаются научить компьютер естественному языку, с другой – использование статистических методов, к которым и относится PLSA (Probabilistic Latent Semantic Analysis – вероятностный латентный семантический анализ).

Первоначально было введено понятие модели векторного пространства [2], в котором любой документ представлялся как вектор частот появления в нем определенных терминов. В этом подходе отношения между документами и терминами задавались в виде матрицы смежности $\boldsymbol{A}$, элементом w_{ij} которой является частота появления термина t_j в документе d_i.

Обозначим через m количество проиндексированных терминов в коллекции документов d, а через n – количество самих документов. В общем случае элементом w_{ij} матрицы $\boldsymbol{A}$ является некоторый вес, поставленный в соответствие паре «документ–термин» (d_i, t_j). После того как все веса заданы, матрица $\boldsymbol{A}$ становится отображением коллекции документов в векторном гиперпространстве. Таким образом, каждый документ можно представить как вектор весов терминов:

$$\boldsymbol{A} = \begin{pmatrix} w_{11} & \bullet & \bullet & w_{1n} \\ \bullet & \bullet & & \bullet \\ \bullet & & \bullet & \bullet \\ w_{m1} & \bullet & \bullet & w_{mn} \end{pmatrix} \equiv (d_1 \quad \bullet \quad \bullet \quad d_n) \equiv \begin{pmatrix} t_1 \\ \bullet \\ \bullet \\ t_m \end{pmatrix}. \quad (1)$$

*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 гг.», № 2011-1.9-519-005.

Подход LSA (Latent Semantic Analysis – латентный семантический анализ), предложенный в работе [3], заключается в отображении документа в латентное семантическое пространство, которое несет в себе основную смысловую нагрузку. Цель такого отображения состоит в том, чтобы отразить скрытую (латентную) связь между терминами и документами, что достигается использованием сингулярного (SVD) разложения матрицы $\boldsymbol{A}$. Оценка схожести документов формируется по близости расположения точек латентного семантического пространства.

В основе методологии PLSA лежит идея, предложенная в LSA и описанная в [4], когда в латентное семантическое пространство вводятся понятия латентного класса:

$$z \in \boldsymbol{Z} = \{z_1, \ldots, z_k\},$$

условных вероятностей среди документов:

$$d \in \boldsymbol{D} = \{d_1, \ldots, d_k\},$$

и терминов

$$w \in \boldsymbol{W} = \{w_1, \ldots, w_k\},$$

а также предполагается, что распределение слов, принадлежащих данному классу, не зависит от документа и пары наблюдений «документ–термин» (d, w) не связаны между собой.

Распределение терминов в документе $P(w|d)$ задается выпуклой комбинацией факторов $P(w|z)$ и $P(z|d)$:

$$P(w \mid d) = \sum_{z \in Z} P(w \mid z) P(z \mid d). \quad (2)$$

Совместная вероятность документа и термина определяется соотношением

$$P(d, w) = P(d) P(w \mid d) = \sum_{z \in Z} P(z) P(d \mid z) P(w \mid z). \quad (3)$$

Используя алгоритм максимизации математического ожидания (Expectation-Maximization, EM Algorithm), который состоит из этапов E и M, можно оценить вероятности $P(w|z)$ и $P(z|d)$, максимизируя логарифмическую функцию правдоподобия:

$$L = \sum_{d \in D} \sum_{w \in W} n(d, w) \log P(d, w), \quad (4)$$

где $n(d, w)$ – частота (количество) появлений термина w в документе d.

Вероятность того что появление термина w в документе d объясняется принадлежностью их к классу z, на этапе E оценивается следующим образом:

$$P(z \mid d, w) = \frac{P(z) P(d \mid z) P(w \mid z)}{\sum_{z' \in Z} P(z') P(d \mid z') P(w \mid z')}. \quad (5)$$

На этапе M происходит переоценка вероятностей:

$$P(w \mid z) = \frac{\sum_{d} n(d, w) P(z \mid d, w)}{\sum_{w'} \sum_{d} n(d, w') P(z \mid d, w')},$$

$$P(d \mid z) = \frac{\sum_{w} n(d, w) P(z \mid d, w)}{\sum_{d', w} n(d', w) P(z \mid d', w)}, \quad (6)$$

$$P(z \mid d, w) = \frac{\sum_{d, w} n(d, w) P(z \mid d, w)}{\sum_{d, w} n(d, w)}.$$

В работе [1] Т. Хофман предложил обобщенную модель для оценивания условной вероятности, которую он назвал *ослабленной процедурой максимизации математического .ожидания* (Tempered Expectation Maximization, TEM). При этом на этапе E в оценку условной вероятности вносится регуляризационный параметр β:

$$P_\beta(z \mid d, w) = \frac{P(z)\left[P(d \mid z) P(w \mid z)\right]^\beta}{\sum_{z' \in Z} P(z')\left[P(d \mid z') P(w \mid z')\right]^\beta}. \quad (7)$$

Согласно (2), любая условная вероятность $P(w|d)$ может быть аппроксимирована полиномом, представляющим собой выпуклую комбинацию условных вероятностей $P(w|z)$. Геометрической интерпретацией весовых коэффициентов $P(z\ d)$ являются координаты документа в подпространстве, определяемом как латентное семантическое пространство [1].

Модель (профиль) пользователя. Рассмотрим схему моделирования интересов пользователя, основанную на инициализации начального профиля и его последовательной корректировке в процессе работы.

Документы могут быть представлены как векторы латентного семантического пространства таким образом, как это показано выше. Однако для того чтобы следить и непрерывно анализировать возможные изменения интересов пользователя, необходимо ввести понятие временного измерения в латентном семантическом пространстве, рассматривая уже не само латентное семантическое пространство, а его модификацию – временное латентное семантическое пространство. Каждое измерение такого векторного пространства (за исключением временного, которое полагается равным нулю) представляет собой условные вероятности при заданном классе $P(\bullet|z)$, а документы являются векторами с весовыми коэффициентами (координатами) $P(z|d)$.

Запросы, как и сами документы, также могут быть представлены в виде векторов во временном латентном семантическом пространстве. Кроме весов $P(z|Q)$ у них имеется дополнительное (временное) измерение (текущий вес), первоначально равный некоторой положительной величине, уменьшающейся с течением времени, исходя из предположения о падении интереса пользователя к определенной тематике при отсутствии ее фигурирования в запросах продолжительное время. Если пользователь инициирует запрос, связанный с категорией из его текущего профиля, то вес данной категории может быть либо стабилизирован на некоторое время, либо увеличен.

Согласно геометрии латентного семантического пространства, запрос, состоящий из терминов, проецируется в латентное семантическое пространство [4]. Таким образом, гиперповерхность S_i, образованная запросом Q_i, является пересечением вероятностных поверхностей всех классов, которые введены

в латентное семантическое пространство и в которых с определенной вероятностью фигурирует данный термин:

$$S_i = \bigcap_k H_{ki}.$$

Алгоритм адаптивной коррекции профиля пользователя основан на неявной обратной связи с пользователем, которая реализуется исходя из истории его запросов. На вход алгоритма поступает запрос пользователя, на выход – одна или более троек (триплетов) вида (C_i, W_i, α_i), где C_i – категория интересов; W_i – текущий вес; α_i – уровень изменчивости (смысл данной величины состоит в том, чтобы отразить, насколько изменяются интересы пользователя в рамках текущего запроса по отношению к прошлым запросам).

Итак, профиль пользователя представляет собой набор троек, организованный таким образом, что интересы пользователя разделены на два типа: краткосрочные (краткосрочный профиль) и долгосрочные (долгосрочный профиль). Как правило, емкость долгосрочного профиля больше емкости краткосрочного. Структуру профиля можно представить таблицей (см. рисунок). При этом считается, что тройки, у которых величина текущего веса положительная, относятся к краткосрочному профилю, а если величина этого веса отрицательная, то к долгосрочному профилю. При этом для троек, находящихся в краткосрочном профиле, текущий вес уменьшается линейно, а для троек, находящихся в долгосрочном профиле, снижение весов экспоненциально.

Кино	Музыка	Квантовая физика	Спорт	Категория
95	85	35	70	Текущий вес
0,60	0,45	0,20	0,15	Уровень изменчивости

Краткосрочный профиль пользователя

Формально профиль в текущий момент i описывается следующим образом:

$$Pr_i = \{(C_j, W_j, \alpha_j)_i, j=1, k\}. \quad (8)$$

При этом

$$Pr_i = PrR_i \cup PrL_i, \quad (9)$$

где $PrR_i = \{(C_j, W_j, \alpha_j)_i \mid \forall\ W_j \geq 0, j=1, k\}$ – краткосрочный профиль; $PrL_i = \{(C_j, W_j, \alpha_j)_i \mid \forall\ W_j < 0, j=1, k\}$ – долгосрочный профиль.

Уровень изменчивости α_i рассчитывается как близость двух последовательных запросов Q_i и Q_{i-1}, представленных в пространстве частот их терминов:

$$\alpha_i = \frac{\sum_w \tilde{n}(Q_i, w)\tilde{n}(Q_{i-1}, w)}{\sqrt{\sum_{w'} \tilde{n}(Q_i, w)^2 \sum_d \tilde{n}(Q_{i-1}, w)^2}}, \quad (10)$$

где $\tilde{n}(Q_i, w)$ – взвешенные частоты терминов.

Алгоритм непрерывной корректировки профиля пользователя. Данный алгоритм в своей работе использует хранилище предыдущих запросов пользователя. В текущий момент времени i пользователь вводит новый запрос, который после соответствующей обработки помещается в хранилище запросов. Обновленное (или дополненное) в момент времени i текущим запросом хранилище запросов будем обозначать Q_i .

Перед тем как передать запрос для работы алгоритму, этот запрос обрабатывается с целью выделения ключевых терминов. Далее производится пересчет взвешенных частот терминов в хранилище запросов Q_i с учетом нового запроса. Когда пользователь вводит очередной запрос, его ключевым словам (терминам) назначаются наибольшие веса. При поступлении запроса в хранилище выполняется проверка на наличие в нем терминов, присутствующих в данном запросе. Если термин встречается впервые, то при его занесении в хранилище вес остается без изменений; если же такой термин в хранилище уже существует (это означает, что пользователь когда-то уже использовал запрос, включающий в себя данный термин), то его весовой коэффициент пересчитывается. В результате происходит нормирование весовых коэффициентов. Категории интересов C_i для включения в текущий профиль пользователя извлекаются из хранилища с использованием методологии PLSA. Алгоритм непрерывной корректировки профиля пользователя состоит из 11 шагов и работает следующим образом.

Шаг 1. Инициализировать хранилище запросов $Q_i = \{w_{1i}, w_{2i}, \ldots, w_{ki}\}$, где w_{ki} – термины хранилища запросов, $k = 1, \ldots, M$.

Шаг 2. Выделить набор ключевых терминов текущего запроса.

Шаг 3. Скорректировать весовые коэффициенты терминов и произвести их нормировку с учетом нового запроса.

Шаг 4. Рассчитать уровень изменчивости α_i.

Шаг 5. Рассчитать условные вероятности классов, используя процедуру ТЕМ:

$$P(z \mid Q_i) = P(w_{ki}) P_\beta(z \mid Q_i, w_{ki}) =$$

$$= \sum_{w_{ki}} P(w_{ki}) \frac{P(z)\left[P(Q_i \mid z) P(w_{ki} \mid z)\right]^\beta}{\sum_{z'} P(z')\left[P(Q_i \mid z') P(w_{ki} \mid z')\right]^\beta}.$$

Шаг 6. Рассчитать вероятность категории C_i для заданного класса латентного семантического пространства:

$$P(C_i \mid z) = \frac{\sum_{Q_i} n(Q_i, C_i,) P_\beta(z \mid Q_i, C_i)}{\sum_{C_i', Q_i'} n(C_i', Q_i) P_\beta(z \mid C_i', Q_i)}.$$

Шаг 7. Рассчитать вероятность включения категории C_i для текущего состояния хранилища запросов Q_i:

$$P(C_i \mid Q_i) = \sum_{z \in Z} P(C_i \mid z) P(z \mid Q_i).$$

Шаг 8. Занести категорию в профиль пользователя, включив соответствующую тройку (C_i, W_i, α_i) в профиль согласно схеме (см. рисунок).

Шаг 9. Если уровень изменчивости $\alpha_i > \alpha_0$, где α_0 – заданная величина, то увеличить текущий вес категории C_i на величину ΔW_i:

$$W_i = W_i + \Delta W_i .$$

Шаг 10. Отсортировать последовательность троек (C_i, W_i, α_i) в профиле по порядку убывания веса W_i .

Шаг 11. Сохранить получившийся профиль.

Таким образом, в данной статье был рассмотрен алгоритм непрерывной корректировки модели (профиля) пользователя. Для успешного построения алгоритма предложена схема организации профиля пользователя в виде множества троек вида (категория интересов C_i, текущий вес категории w_i, уровень изменчивости α_i). При этом профиль делится на две группы (два подпрофиля): краткосрочный и долгосрочный – для учета краткосрочных и долгосрочных интересов пользователя (в том числе неявных). Кроме того, было введено понятие временного измерения в латентном семантическом пространстве, что позволило адаптировать методологию PLSA для непрерывной оценки изменений интересов пользователя.

Применение предложенного алгоритма для подстройки модели в процессе ее работы с использованием неявной обратной связи, приближает нас к созданию высококачественных и эффективных поисковых систем с персонализированным интерфейсом.

Библиографические ссылки

1. Hoffman T. Unsupervised Learning by Probabilistic Latent Semantic Analysis // Machine Learning. 2008. Vol. 42. P. 177–196.

2. Salton G., McGrill M. J. Introduction to Modern Information Retrieval. New York : McGraw-Hill, 1993.

3. Indexing by Latent Semantic Analysis / S. Deerwes, S. Dumasis, G. Furnas et al. // J. of the Amer. Soc. for Inform. Science. 1990. Vol. 41. P. 391–407.

4. Hoffman T. Probabilistic Latent Semantic Indexing // Proc. of the 22nd Annu. Intern. ACM SIGIR Conf. on Research and Development in Inform. Retrieval. Berkeley, Calif., 2009. P. 50–57.

M. V. Karasyova

APPLICATION OF PLSA METHODOLOGY FOR ADAPTIVE CORRECTION OF USER MODEL

The paper considers the algorithm of the continuous model (profile) correction. The initial data are the initial profile and the previous inquiry history. The PLSA (Probabilistic Latent Semantic Analysis) methodology is used in the algorithm. To achieve the object in view, the term temporary latent semantic space is introduced.

Keywords: probabilistic latent semantic analysis, user model, user profile.

УДК 519.254

В. С. Кедрин, О. В. Кузьмин

ВЫДЕЛЕНИЕ ОСЦИЛЛИРУЮЩИХ И ТРЕНДОВЫХ КОМПОНЕНТ НА БАЗЕ КРИТЕРИАЛЬНОЙ МОДИФИКАЦИИ СИНГУЛЯРНОГО АНАЛИЗА

Рассматривается методика выделения осциллирующих колебательных и трендовых составляющих при анализе сложных нестационарных процессов, протекающих в реальных сложных системах.

Ключевые слова: нестационарная система, временной ряд, сингулярное разложение, графические критерии качества, осциллирующие компоненты, трендовые компоненты.

Одним из интенсивно развивающихся теоретических подходов к моделированию сложных процессов является использование непараметрических моделей динамических систем в виде набора элементарных характеристик (временных выборок), позволяющих по экспериментальным данным входа-выхода выявить динамические свойства и оценить состояние исследуемой системы. В этой связи особо актуальными становятся исследования, посвященные анализу факторов, влияющих на состоятельность практических рекомендаций в области идентификации и синтеза цифровых непараметрических моделей систем, в которых протекают процессы с ярко выраженными нестационарными свойствами. Примерами таких систем являются современные электроэнергетические системы и системы товарных и фондовых рынков.

В настоящее время существует несколько подходов к решению проблем, связанных с анализом нестационарной динамики.

Один из этих подходов основан на методах, заимствованных из нелинейной динамики [1–5]. Однако данный подход является ограниченным, так как он подразумевает постоянство оператора эволюции системы, в силу чего нестационарность, вызванная