Информационные ресурсы и технологии

КОВАЛЕВ Игорь Владимирович *– доктор технических наук, профессорКрасноярского государственного технического университета*

СЛОБОДИН Михаил Юрьевич – *кандидат технических наук**,** докторант НИИ Систем управления, волновых процессов и технологий*

КУСТОВ Денис Александрович *– аспирант Сибирского государственного аэрокосмического университета*

ИНТЕЛЛЕКТУАЛИЗАЦИЯ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ В КОРПОРАТИВНЫХ СИСТЕМАХ

В данной статье авторами рассматриваются новые подходы к интеллектуализации информационных технологий в корпоративных системах (КИС). К настоящему времени, когда многие технические вопросы (способы хранения данных, распределения ресурсов между удаленными терминалами, способы передачи данных и т.п.) создания КИС уже достаточно проработаны, на передний план выходит интеллектуализация таких систем. Одной из проблем, с которой могут столкнуться разработчики КИС в современных условиях, является языковое взаимодействие специалистов мультинациональной компании. В связи с этим предлагается еще на этапе проектирования КИС интегрировать в нее аппарат программно-алгоритмической поддержки мультилингвистической адаптивно-обучающей технологии. В основе данной технологии лежит использование частотных терминологических словарей.

Поскольку в настоящее время процесс создания сложных компьютерных систем является модульным, то техническая сторона реализации предлагаемых подходов, очевидно, может быть реализована в виде дополнительных (опциональных) модулей (рис. 1). Эти модули могут отсутствовать в стандартной конфигурации КИС и включаться разработчиком (или компанией интегратором) по желанию заказчика.

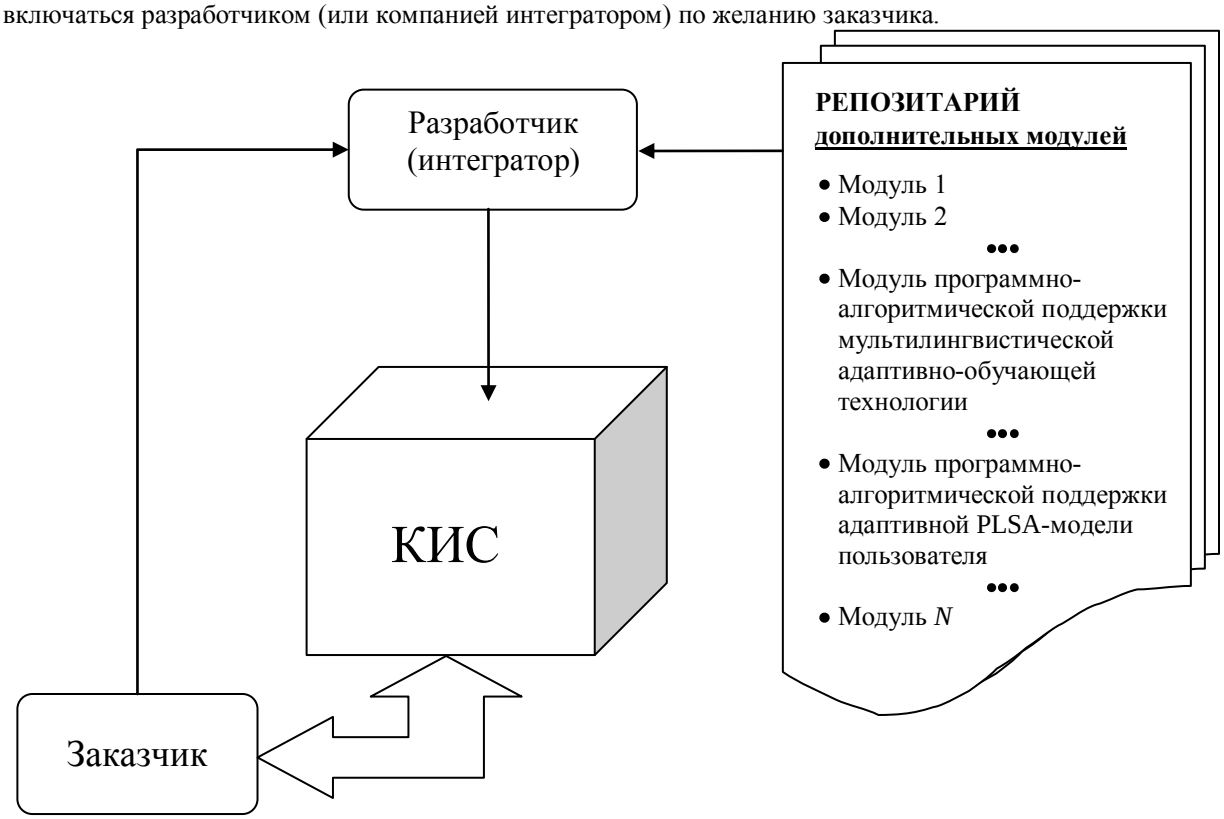

Рис. 1. Модульный подход при разработке современной КИС

Мультилингвистическая адаптивно-обучающая технология

Следует отметить, что в контексте создания и поддержания КИС, в большей мере, предполагается взаимодействие специалистов одного уровня, т.е. работающих в одной области. Например, конструкторы, технологи или экономисты. В первую очередь на возможность взаимодействия специалистов компании, принадлежащих разным языковым группам, влияет знание специализированной терминологии.

В данной статье кратко затронуты вопросы межъязыкового взаимодействия специалистов различных языковых групп. За более подробной информацией можно обратиться к [1].

Применение информационной мультилингвистической обучающей технологии способствует более эффективному использованию алгоритма обучения в рамках создаваемой КИС, позволяя, кроме того, интенсивно пополнять профессионально-ориентированный активный словарь одновременно нескольких языков.

Необходимо выделить ряд основных достоинств такого подхода к организации компьютерной системы:

- учитывается фактор частотности слов (т.е. для наиболее быстрого и эффективного обучения заучивается не весь словарь, а, в первую очередь, та его часть, которая чаще других употребляется в тексте);

- учитывается индивидуальная специфика памяти человека (т.е. повторяются не все слова подряд, а те, которые хуже запоминаются или быстрее забываются);

- промежутки между сеансами обучения произвольны, что особенно важно для применения системы в реальной обстановке;

- учитывается отличие объема очередной порции обучающей информации на каждом сеансе от общего ее объема;

- мультилингвистичность обеспечивает генерацию ассоциативного поля вокруг запоминаемых понятий;

- учитывается такое важное свойство человеческой памяти, как уменьшение скорости забывания обучающей информации по мере ее повторения;

- специфичной для использования в КИС является возможность обучения специализированной терминологии без отрыва работника от выполнения основных функциональных обязанностей.

Следует учитывать, что модель обучаемого, которая используется в реализации алгоритма, является результатом исследований в области психологии. Это экспоненциальная зависимость вероятности незнания лексической единицы от скорости ее забывания, степени ее связи с иностранными значениями и времени с момента ее последнего заучивания.

Для достижения целей, поставленных при формировании информационно-терминологической базы мультилингвистической (МЛ) обучающей технологии, в первую очередь, необходимо разработать электронные частотные мультилингвистические словари для различных областей знаний. Так как эти словари представляют собой словарную базу разрабатываемых компьютерных систем изучения терминологической лексики иностранных языков, то правильность их составления (выбор терминов) и определение частотных характеристик терминов существенно влияет на эффективность работы системы обучения. Указанные характеристики входят в состав модели обучения и непосредственно влияют на качество процесса обучения.

В рамках средств структурного анализа, эффективно применяемых в информационных системах, важное место занимают словари данных, организация которых в МЛ-технологии также имеет свои особенности из-за мультилингвистического характера базисных компонентов. Однако, следуя методологии структурного анализа и применяя DFD (Data Flow Diagrams), можно показать внешние по отношению к системе источники и стоки (адресаты) данных, идентифицировать логические функции (процессы) и группы элементов данных, связывающие одну функцию с другой (потоки), а также идентифицировать хранилища (накопители) данных, к которым осуществляется доступ при реализации функций МЛ-технологии. Структуры потоков мультилингвистических данных и определения их базисных компонент хранятся и анализируются в МЛ-словаре данных, что требует его специальной организации.

Как правило, словарь данных представляет собой определенным образом организованный список всех элементов данных системы с их точными определениями (для МЛ-словаря, следовательно, определения должны быть многоязычными), что дает возможность различным категориям пользователей (от системного аналитика, алгоритмиста до специалиста по компьютерной лингводидактике) иметь общее понимание всех входных и выходных потоков и компонент хранилищ. Определения элементов данных в словаре осуществляются следующими видами описаний:

- описанием значений потоков и хранилищ, изображенных на DFD;

- описанием композиции агрегатов данных, движущихся вдоль потоков, то есть, комплексных данных, которые могут расчленяться на элементарные символы (например, для МЛ-словаря любой термин или базисный информационный компонент может содержать на уровне сеанса обучения такие символы, как терминологическое множество других языков, множество частотных характеристик терминов и т.д.);

- описанием композиции групповых данных в хранилище (например, на каждом сеансе базисному информационному компоненту МЛ-словаря ставится в соответствие композиция альтернативных многоязычных терминов и многоязычных подсказок; мощность группы данных определяется алгоритмами, реализуемыми в МЛ-технологии обучения);

- специфицированием значений и областей действия элементарных фрагментов информации в потоках данных и хранилищ; описанием деталей отношений между хранилищами.

Современные программно-алгоритмические средства разработки компьютерных систем требуют применения оригинальных методик при формировании информационной модели данных, используемых при обучении. Разрабатываемые авторами методики базируются на основных идеях структурного системного анализа и на структурных методологиях, относящихся к классу методологий, ориентированных на данные. С позиций ориентированных на данные методологий вход и выход информационной модели являются наиболее важными, структуры данных (а не потоки данных) определяются первыми, а компоненты обучающих процедур технологии строятся как производные от структур данных.

Так терминологическое множество, соответствующее базисному информационному компоненту мультилингвистической обучающей технологии, может быть описано следующим образом:

МЛ-компонент = {*термин яз_1, термин яз_2,..., термин яз_N, частота яз_1, частота яз_2,..., частота яз_N*}.

DSSD использует аналогичную нотацию, а именно множественную скобку (рис. 2).

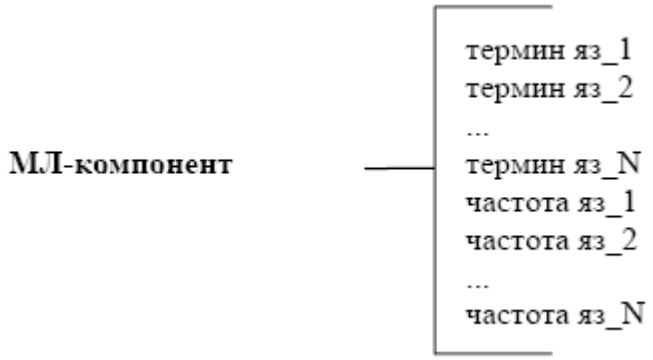

Рис. 2. Описание базисного информационного компонента с использованием нотации DSSD

В качестве модели обучаемого предлагается модификация адаптивной модели Л.А.Растригина применительно к мультилингвистической адаптивно-обучающей технологии. Как и в работах Растригина Л.А. используется широко распространенная аппроксимация процессов обучения монотонной экспоненциальной кривой, однако, учитывая, что во многих экспериментах выявлено несовпадение реального процесса обучения с экспоненциальной кривой, система обучения предполагает расширение базиса моделей и возможность работы с экспериментальными данными для построения реальных зависимостей.

Состояние обучаемого на *n*-м сеансе описывается вектором вероятностей незнания каждого из элементов обучающей информации (ОИ), где вероятность незнания *i*-го элемента определяется в каждый *n*-й момент времени t_n. Вероятности незнания элементов ОИ изменяются по правилу, определяемом экспоненциальным законом, учитывающим скорость забывания *i*-го элемента ОИ на *n*-м сеансе. Учитываются также и его связи с элементами ранее изучавшихся иностранных терминологий. Скорость забывания каждого элемента уменьшается, если этот элемент выдается обучаемому для запоминания и не изменяется, если он не заучивается [1].

Начальное значение скорости забывания оценивается методом максимального правдоподобия [3]. На основе указанного метода также оцениваются и параметры коррекции скоростей забывания, характеризующие индивидуальные особенности памяти обучаемого. Критерий качества обучения Q_n, учитывающий ответы обучаемого на тесты, характеризует уровень обученности. Для рассматриваемой задачи данный уровень характеризуется вероятностью незнания элемента ОИ, наугад выбранного из текста, с учетом вероятности незнания *i*-го элемента ОИ, а также относительной частоты, выражающей долю лексической единицы в тексте, подвергшемуся статистической обработке при составлении частотного словаря. При статистической обработке используется абсолютная частота появления лексической единицы в тексте, частоты из мультилингвистического словаря соответственно при английском, немецком и русском слове (в случае трехъязычной среды информационного взаимодействия), а также объем текста при составлении частотного словаря.

Методология PLSA в области извлечения информации

Проблема извлечения информации в информационных ресурсах (к которым относятся различные хранилища, репозитарии и т.п.) получила новое развитие в связи с появлением всемирной сети Интернет. В настоящее время каждый пользователь, имеющий доступ к интернету, имеет доступ ко всем источникам информации, представленным в ней. Казалось бы, что теперь со своевременным получением необходимой информации по интересующей тематике не должно возникнуть больших проблем: ввел запрос поисковой машине и жди ответа в виде набора ссылок на интересующие документы. Однако на деле оказалось, что качество поиска информации при всей ее доступности очень низкое. В поисковых машинах (иначе их называют поисковыми сервисами) отсутствуют эффективные алгоритмы поиска релевантной информации (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 ни какой связи с сутью запроса.

Среди исследователей можно выделить приверженцев двух идей. С одной стороны – традиционная лингвистическая школа, пытающаяся научить компьютер естественному языку. С другой, сообщество, ориентированное на использование статистических методов. Подход PLSA относится ко второй группе.

Первоначально было введено понятие модели векторного пространства [7]. При этом любой документ представлялся как вектор частот появления определенных терминов в нем. В этом подходе отношения между документами и терминами представлялись 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:

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

Подход LSA (Latent Semantic Analysis – Латентный Семантический Анализ), предложенный в 1990 в работе [8], заключается в отображении документа в латентное семантическое пространство. Основная цель – отразить скрытую (латентную) связь между терминами и документами. Это достигается использованием сингулярного разложения (***SVD***-разложение) матрицы ***A***. Предполагается, что такое пространство несет в себе основную смысловую нагрузку. Оценка схожести документов формируется по близости расположения точек латентного семантического пространства.

В основе методологии PLSA лежит идея, предложенная в LSA и описанная в [6]. В рамках PLSA на латентном семантическом пространстве вводится понятие латентного класса $z \in \boldsymbol{Z}$. Также рассматриваются условные вероятности среди документов $d \in \boldsymbol{D}$ и терминов $w \in \boldsymbol{W}$.

Далее предполагается, что распределение слов, принадлежащих данному классу, не зависит от документа и пары наблюдений документ-термин (d, w) – независимы.

Распределение терминов в документе $P(w|d)$ задается выпуклой комбинацией факторов $P(w|z)$ и $P(z|d)$. Определяется сумма произведений условных вероятностей на заданном классе. Затем определяется совместная вероятность документа и термина.

Используя алгоритм максимизации математического ожидания (Expectation-Maximization, EM algorithm), который состоит из двух этапов (E и M), оцениваются вероятности $P(w|z)$ и $P(z|d)$, максимизируя логарифмическую функцию правдоподобия с использованием частоты терминов в документе (количество появлений термина w в документе d). При этом оценивается вероятность того, что появление термина w в документе d объясняется принадлежностью их к классу z, на шаге E. На этапе M происходит переоценка вероятностей $P(w|z)$, $P(d|z)$, $P(z|d,w)$ согласно формулам, описанным в [2].

Т. Хофман в работе [5] предложил обобщенную модель для оценивания условной вероятности, которую он назвал ослабленной процедурой максимизации математического ожидания (TEM – Tempered Expectation Maximization). При этом на этапе E в оценку условной вероятности вносится регуляризационный параметр β.

Согласно [5] любая условная вероятность $P(w|d)$ может быть аппроксимирована полиномом, представляющим собой выпуклую комбинацию условных вероятностей $P(w|z)$. Геометрическая интерпретация весовых коэффициентов $P(z|d)$ – координаты документа в подпространстве, определяемом как латентное семантическое пространство.

Модель пользователя (профиль пользователя)

Авторами предлагается новая схема моделирования интересов пользователя, основанная на инициализации начального профиля и его последовательной корректировке в процессе работы.

Документы могут быть представлены как векторы латентного семантического пространства так, как это показано выше (за более подробным описанием обращайтесь к [5,6]). Для того чтобы следить и непрерывно анализировать возможные изменения интересов пользователя, предлагается ввести понятие *временного* измерения в латентном семантическом пространстве, тем самым, рассматривая уже не само латентное семантическое пространство, а его модификацию – временное латентное семантическое пространство. Каждое измерение (за исключением временного) такого векторного пространства, аналогично описанному в предыдущем разделе, представляет собой условные вероятности при заданном классе $P(\bullet|\ z)$. Документы представляют собой векторы с весовыми коэффициентами (координатами) $P(z|d)$, временное измерение полагаем равным нулю.

Запросы, равно как и сами документы, могут быть представлены в виде векторов во временном латентном семантическом пространстве. Кроме весов у них есть 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 определенной категорией из его текущего профиля, то вес данной категории может быть либо стабилизирован на определенное время, либо увеличен.

Согласно геометрии рассмотренного в латентного семантического пространства запрос, состоящий из терминов, проецируется в латентное семантическое пространство [6]. Таким образом, гиперповерхность S_i, образованная запросом Q_i является пересечением вероятностных поверхностей всех классов, введенных на латентном семантическом пространстве, в которых с определенной вероятностью фигурирует данный термин.

Алгоритм адаптивной коррекции профиля пользователя основан на неявной обратной связи с пользователем, которая реализуется на основе истории его запросов. На вход алгоритма поступает запрос пользователя, на выходе – 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. При этом он организован таким образом, что интересы пользователя были разделены на два типа: краткосрочные (краткосрочный профиль) и долгосрочные

(долгосрочный профиль). Как правило, емкость долгосрочного профиля больше емкости краткосрочного. Структуру профиля можно представить следующей таблицей.

Краткосрочный профиль пользователя

Категория	ино	Музыка	Квантовая физика	порт
Текущий вес	5	85	35	0
Уровень изменчивости	.60	0.45	0.20	.15

При этом считается, что тройки, в которых величина текущего веса положительная, относятся к краткосрочному профилю, если вес отрицательный – то к долгосрочному профилю. При этом для троек, находящихся в краткосрочном профиле, текущий вес уменьшается линейно, тогда как для троек, находящихся в долгосрочном профиле снижение весов – экспоненциально.

Формально профиль в текущий момент i может быть описан как PrR_i – краткосрочный профиль и PrL_i – долгосрочный профиль.

Уровень изменчивости профиля (α_i) рассчитывается как близость двух последовательных запросов Q_i и Q_{i-1} представленных в пространстве частот их терминов с учетом взвешенных частот терминов.

Алгоритм непрерывной корректировки профиля пользователя

При использовании данного алгоритма предполагается, что существует некоторое 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, производится его обработка на предмет выделения ключевых терминов.

Далее производится пересчет взвешенных частот терминов в хранилище запросов Q_i с учетом нового запроса. Когда пользователь вводит очередной запрос, ключевым словам (терминам) данного запроса назначаются наибольшие веса. При поступлении запроса в хранилище запросов происходит проверка на наличие в хранилище терминов, присущих текущему запросу. Если термин встречается впервые, то при его занесении в хранилище вес остается без изменений, если же такой термин уже существует в хранилище (это означает, что пользователь уже когда-то использовал запрос, включающий данный термин), то производится пересчет весового коэффициента данного термина. В конечном счете, происходит нормирование весовых коэффициентов.

Категории интересов C_i для включения в текущий профиль извлекаются из хранилища посредством использования методологии PLSA, описанной выше.

Далее представлен пошаговый алгоритм непрерывной корректировки профиля пользователя.

1. Инициализировать хранилище запросов $Q_i = \{w_{1i}, w_{2i}, \dots , w_{ki}\}$, где w_{ki} – термины хранилища запросов, $k = 1 \dots M$.
2. Выделить набор ключевых терминов текущего запроса.
3. Скорректировать весовые коэффициенты терминов и произвести их нормировку с учетом нового запроса.
4. Рассчитать уровень изменчивости α_i.
5. Рассчитать условные вероятности классов $P(z|Q_i)$, используя процедуру TEM [2].
6. Рассчитать вероятность $P(Ci|z)$ категории C_i для заданного класса латентного семантического пространства.
7. Рассчитать вероятность включения категории C_i для текущего состояния хранилища запросов Q_i .
8. Занести категорию в профиль пользователя. Для этого включить соответствующую тройку (C_i, W_i, α_i) в профиль.
9. Если уровень изменчивости $\alpha_i > \alpha_0$ (где α_0 заданная величина), то увеличить текущий вес категории C_i на величину ΔW_i : $W_i = W_i + \Delta W_i$.
10. Отсортировать последовательность троек (C_i, W_i, α_i) в профиле по порядку убывания веса W_i.
11. Сохранить получившийся профиль как текущий.

Экспериментальные исследования

Эффективность методов информационного поиска оценивается на тестовых наборах данных. В течение последнего десятка лет был создан ряд стандартных тестовых наборов данных, которые в настоящее время повсеместно используются для проведения экспериментов в области информационного поиска.

Для исследования описанного алгоритма использовалось 4 набора документов:

- MED – 1033 документа из Национальной медицинской библиотеки;
- CRAN – 1400 документов по авиационной тематике;
- CACM – 3204 статьи из журналов CACM (Communications of the Association for Computing Machinery – Средства Связи Ассоциации Компьютерной Технологии);
- CISI – 1460 документов из научной библиотеки.

В качестве критериев качества реализуемых подходов к решению задач информационного поиска используются следующие: точность (*Precision*) и полнота (*Recall*) ответа [7]. Обозначим через ***C*** – коллекцию документов, в которой осуществляется поиск, ***A*** – множество документов-ответов на запрос, ***R*** – множество истинно релевантных документов.

Результаты экспериментальных исследований представлены на рисунке 3.

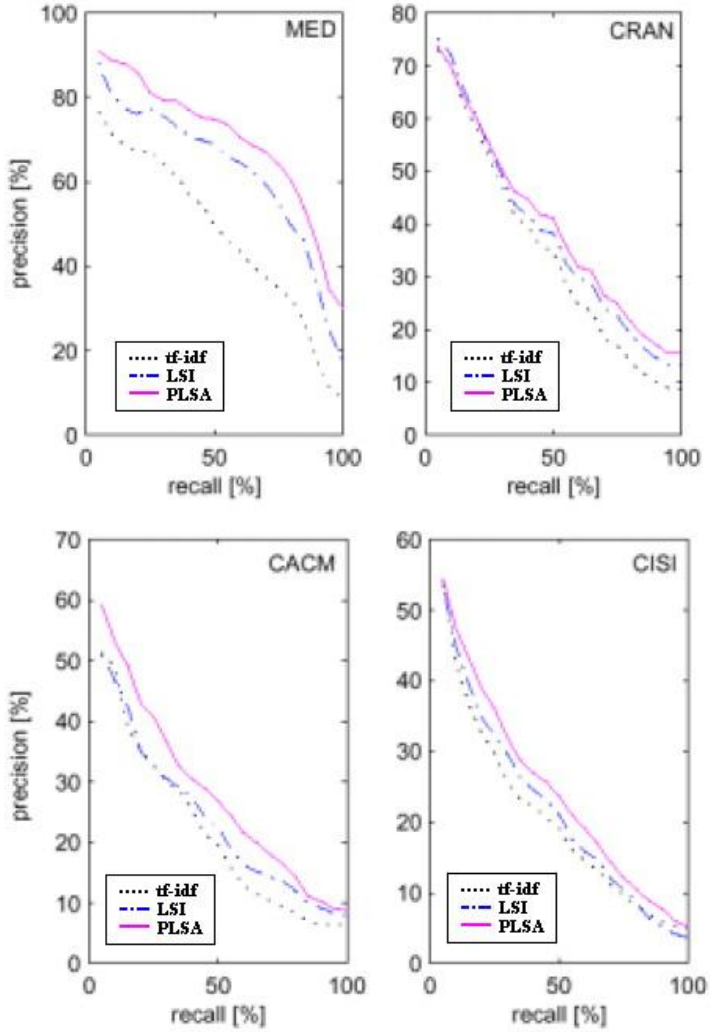

Рис. 3. Результаты сравнения подходов

Из представленных графиков очевидно, что разработанный подход дает увеличение качества поиска на всех рассмотренных коллекциях документов (см. таблицу ниже).

	M ED	**C RAN**	**C ACM**	**ISI**
PLSA vs tf-idf	5 -20%	1 -10%	5 -10%	-5%
PLSA vs LSI	3 -8%	0 -2%	1 -5%	-5%

Итак, для успешной реализации алгоритма непрерывной корректировки модели (профиля) пользователя предложена схема организации профиля пользователя в виде множества троек вида: *категория интересов*, *текущий вес категории*, *уровень изменчивости*. Для использования в корпоративных информационных системах профиль пользователя делится на две группы (два подпрофиля): краткосрочный и долгосрочный для учета краткосрочных и долгосрочных интересов пользователя при формировании информационно-терминологического базиса. Введенное авторами понятие временного измерения в латентном семантическом пространстве позволило адаптировать методологию PLSA для непрерывной оценки изменений интересов пользователя. Таким образом, учет неявных интересов пользователя, отраженных в его модели, а также использование предложенного алгоритма для подстройки модели в процессе работы с использованием неявной обратной связи с пользователем приближает нас к созданию высококачественных информационных систем с элементами интеллектуализации и персонализированным интерфейсом.

Литература

1. И.В. Ковалев, М.В. Карасева, Е.А. Суздалева Системные аспекты организации и применения мультилингвистической адаптивно-обучающей технологии//Educational Technology & Society.- 2002. - 5(2) .

2. Ковалев И.В., Кустов Д.В. PLSA-адаптация модели пользователя в открытой информационно-образовательной среде // Телекоммуникации и информатизация образования. - 2004.- №6 (25).

3. Ковалев И.В. Системная архитектура мультилингвистической адаптивно-обучающей технологии и современная структурная методология // Телекоммуникации и информатизация образования. - 2002. - №3 (10).

4. Захарушкин В.Ф. Особенности создания информационного обеспечения корпорации // Электронный журнал "Исследовано в России". 2003.

5. Hoffman T. Unsupervised Learning by Probabilistic Latent Semantic Analysis// Machine Learning. - 2001. - Vol. 42.

6. Hoffman T. Probabilistic Latent Semantic Indexing//Proc. Of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 1999.

7. SaltonG., McGrill M.J. Introduction to Modern Information Retrieval. McGraw-Hill, New York, 1993.

8. Deerwester S., Dumasis S., Furnas G., Landauer T., Harshman R. Indexing by Latent Semantic Analysis.// Journal of the American Society for Information Science. - 1990. - Vol. 41.