На правах рукописи

Лесков Виталий Олегович

МУЛЬТИЛИНГВИСТИЧЕСКИЕ СИСТЕМЫ АДАПТИВНОГО ОБУЧЕНИЯ НА БАЗЕ ЛЕКСИЧЕСКИ СВЯЗАННЫХ ИНФОРМАЦИОННЫХ КОМПОНЕНТОВ

05.13.01 – Системный анализ, управление и обработка информации
(по отраслям: информатика, вычислительная техника и управление)

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Красноярск – 2009

Работа выполнена в Федеральном государственном образовательном учреждении высшего профессионального образования «Сибирский федеральный университет», г. Красноярск

Научный руководитель: доктор технических наук, профессор
Ковалев Игорь Владимирович

Официальные оппоненты: доктор технических наук, профессор
Доррер Георгий Алексеевич
кандидат технических наук, доцент
Усачев Александр Владимирович

Ведущая организация: Государственный научно - исследовательский институт информационных технологий и телекоммуникаций «Информика» (г. Москва).

Защита состоится "13" ноября 2009 года в 14:00 часов на заседании диссертационного совета ДМ 212.099.06 при Сибирском федеральном университете по адресу: 660074, г. Красноярск, ул. академика Киренского, 26, УЛК - 115.

С диссертацией можно ознакомиться в библиотеке Сибирского федерального университета по адресу: г. Красноярск, ул. академика Киренского, 26, Г274.

Автореферат разослан "__" октября 2009 года.

Ученый секретарь
диссертационного совета Р.Ю. Царев

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность работы.

Развитие современных технологий, будь-то технологии производства или информационные технологии, неизменно влечет за собой появление новых текстов описательного свойства. Технические документации, учебники, самоучители – это далеко не весь спектр литературы, сопровождающей новое технологическое оборудование, программные продукты, а так же те или иные изменения в технологических процессах.

Огромные объемы и интенсивность появления специальных текстов, необходимость коммуникации с зарубежными партнерами, вынуждают российский рынок труда ужесточать требования, предъявляемые к специалистам тех или иных областей знания.

И если требования к профессиональным навыкам, знанию и опыту кандидата остаются неизменно высокими, то требования к владению иностранными языками с каждым годом все возрастают. При этом от специалиста не требуется глубокого знания того или иного иностранного языка, требования, как правило, ограничиваются предметной областью его профессиональной деятельности.

В рамках решения этой проблемы обучение иностранному языку во многих неязыковых ВУЗах имеет профессиональную ориентацию. Одним из основных моментов в таком обучении является освоение характерной для учебной специальности иностранной терминологии.

Развитие информационных технологий на сегодняшний день позволяет разработать эффективный программно-алгоритмический аппарат для создания компьютерных интерактивных обучающих систем, в том числе, для изучения иностранной терминологической лексики.

Лежащая в основе диссертационного исследования, мультилингвистическая адаптивно-обучающая технология (МЛ-технология) разработана специально для создания подобных систем. Она базируется на адаптивной модели обучаемого Л.А. Растригина и электронных мультилингвистических частотных словарях. К ее достоинствам можно отнести адаптивную подстройку параметров модели обучаемого, генерацию ассоциативных полей между терминами-аналогами различных языков и, как следствие, корректировку скоростей забывания, используемых в рамках обучающего алгоритма. Однако МЛ-технология, как и многие другие при обучении не использует в полной мере лексические зависимости в рамках изучаемого языка, поэтому генерация внутриязыковых ассоциативных полей происходит исключительно на этапе применения знаний на практике и имеет стихийный характер. Если же такая генерация будет происходить организованно и непосредственно в процессе обучения, качество такого процесса, а, следовательно, и обучающих систем в значительной степени возрастет.

Описанные проблемы в сфере образования, на рынке труда, в информационных технологиях определи актуальность работы; необходимость

в построении новой методики обучения, разработки соответствующего алгоритмического аппарата и информационной поддержки дали простор для дальнейшего исследования.

Объект исследования – адаптивные программные системы учебного назначения, структуры данных информационной поддержки обучающих систем и сопутствующие им алгоритмы.

Предмет исследования – программные системы, реализующие мультилингвистическую адаптивно-обучающую технологию, и связанные с ними модели данных, алгоритмы обучения, обработки текстов и формирования частотных словарей.

Целью диссертационной работы является разработка программно-алгоритмических и информационных средств поддержки мультилингвистической адаптивной системы обучения, ориентированной на формирование у обучаемого внутриязыковых ассоциативных связей непосредственно в процессе обучения.

Задачи исследования обусловлены поставленной целью и включают:

– теоретико-информационный анализ методов и алгоритмов мультилингвистической адаптивно-обучающей технологии как сложной проблемно-ориентированной системы управления;

– теоретико-информационный анализ структуры информационной базы мультилингвистической адаптивно-обучающей технологии;

– разработка методики обучения многоязычной иностранной терминологии как совокупности алгоритмов и методов управления сложным обучающим процессом, ориентированным на формирование у обучаемого внутриязыковых ассоциативных связей;

– разработка структуры частотных лексически связанных словарей как средства информационной поддержки адаптивных мультилингвистических систем;

– модификация адаптивного алгоритма обучения, учитывающая особенности лексически связанной структуры информационно-терминологического обеспечения и предложенной методики обучения в целом;

– разработка алгоритмов структурно-параметрического синтеза информационно-терминологического базиса как совокупности лексически связанных компонентов;

– формирование информационно-терминологического базиса предложенной структуры и реализация его в виде электронного двухблочного мультилингвистического лексически связанного словаря на основе разработанных алгоритмов и других компьютерных методов обработки информации.

Методы исследования. При выполнении работы использовались методы системного анализа, оптимизации, методологии структурного анализа, аппарат теории управления, адаптации сложных систем и теории вероятностей.

Новые научные результаты, полученные лично автором.

1. Предложена методика обучения многоязычной иностранной терминологии как совокупность алгоритмов и методов управления сложным обучающим процессом, ориентированным на формирование у обучаемого внутриязыковых ассоциативных связей.
2. Модифицирован алгоритм адаптивного обучения с учетом лексически связанной структуры информационных компонентов терминологического базиса и неоднородности скоростей забывания лексем.
3. Для внедрения предложенной методики в процесс обучения посредством применения интерактивных мультилингвистических адаптивно-обучающих систем разработана структура частотных лексически связанных словарей как средства их информационно-терминологической поддержки.
4. Разработаны, программно реализованы и протестированы алгоритмы структурно-параметрического синтеза информационно-терминологического базиса как совокупности лексически связанных информационных компонентов.

Научная новизна результатов работы.

1. Предложенная методика обучения многоязычной иностранной терминологии является совокупностью новых алгоритмов и методов управления сложным обучающим процессом. В отличие от стандартных методик рассматривает обучающий процесс с точки зрения формирования у обучаемого внутриязыковых ассоциативных связей; отталкивается от новой, разработанной автором, структуры информационно-терминологического базиса.
2. Модифицированный алгоритм адаптивного обучения представляет собой средство поддержки предложенной автором методики обучения, учитывает неоднородность скоростей забывания лексем и поддерживает целостность структуры информационных компонентов терминологического базиса.
3. Предложенная структура информационно-терминологической поддержки разработана на базе частотных словарей, применяемых в МЛ-технологии; изменена структура базисного информационного компонента; выделены наиболее значимые лексемы; учтены лексические связи.
4. Разработанные алгоритмы структурно-параметрического синтеза информационно-терминологического базиса являются новыми, так как формируют базис в виде совокупности лексически связанных информационных компонентов.

Значение для теории. Результаты, полученные при выполнении диссертационной работы, имеют существенное значение для теории построения мультилингвистических адаптивно-обучающих и направлены на повышение качества этих систем, снижения скоростей забывания терминологии и ускорения процессов интеграции полученного знания в практику. Данные результаты являются значимыми для развития модельно-алгоритмического обеспечения систем анализа информации на основе компьютерных методов обработки информации.

Практическая ценность. Разработанная в рамках диссертационного исследования как совокупность алгоритмов и методов управления сложным обучающим процессом методика обучения позволяет уже на этапе освоения иностранной терминологии формировать у обучаемого строго организованные системы внутриязыковых ассоциативных связей. При классическом подходе эти системы формируются стихийно, когда полученные знания применяются на практике; на первых порах они отличаются неустойчивостью связей и относительной нечеткостью структуры, что сказывается на качестве знаний обучаемого, скорости забывания лексем и возникновения путаницы между терминами. Реализация предложенной методики в интерактивных системах обучения, или применение ее на некотором терминологическом базисе позволяет предупредить описанные выше проблемы. В сравнении с «классическим» подходом к обучению, ее применение на практике в конечном итоге дает более качественное, структурированное знание специальной иностранной терминологии: лексемы, выбранные как основные, заучиваются в несколько раз лучше (зависит от силы и количества связей в ЛС-компоненте), а наличие ассоциативных зависимостей существенно замедляет процесс забывания терминологии в целом.

Также, на основе разработанных в диссертации алгоритмов, структур данных построен информационно-терминологический базис (ИТБ) в форме двухблочного трехуровневого электронного англо-немецко-русского частотного словаря по информатике и системному анализу. Данный программный продукт предназначен для изучения специальной терминологии английского и (или) немецкого языка, являясь лексически связанным, поддерживает разработанную методику обучения.

Реализация результатов работы.

Работа выполнялась в рамках ряда проектов аналитической ведомственной целевой программы «Развитие научного потенциала высшей школы» (2007-2008 г.г.), в частности РНП 2.2.2.3.9676 «Модельно-алгоритмическое обеспечение мультилингвистической технологии интерактивного формирования многоязычных информационных ресурсов» и РНП 2.2.2.3.10144 «Программно-информационная технология интерактивного формирования многоязычных частотных словарей терминологической лексики»; а так же по проектам тематического плана СФУ (2008-2012 гг.).

Наиболее важные алгоритмы и структуры данных, разработанные в ходе диссертационной работы, получили программную реализацию и восемь из них зарегистрированы в Отраслевом фонде алгоритмов и программ.

Апробация работы. Основные положения и результаты работы прошли всестороннюю апробацию на всероссийских конференциях, научных семинарах и научно-практических конференциях. В том числе:

– V Всероссийская научно-практическая конференция «Недра Кузбасса. Инновации - 2006», 29 января 2006 года, г. Кемерово;

– Всероссийская научно-практическая конференция РАЕ "Новые информационные технологии и системы", 15-20 декабря 2008 года, г. Москва;

– Всероссийская научно-практическая конференция РАЕ "Исследования в области образования, молодежной политики и социальной политики в сфере образования", 15-20 января 2009 года, г. Москва;

– Всероссийская заочная электронная конференция РАЕ "Современные наукоемкие технологии", 15-20 февраля 2009 года, г. Москва;

– Всероссийская заочная электронная конференция РАЕ "Образовательные технологии", 15-20 марта 2009 года, г. Москва.

Диссертационная работа в целом обсуждалась на научных семинарах кафедры информатики Сибирского федерального университета в 2007-2009 гг.

Публикации. Основные результаты диссертационной работы опубликованы в 15 работах автора, 8 из них опубликовано в ведущих рецензируемых научных журналах и изданиях, рекомендуемых ВАК РФ для опубликования основных научных результатов диссертационных исследований. Полный список публикаций помещен в конце автореферата.

Структура и объем работы. Диссертационная работа состоит из введения, четырех глав, заключения и списка использованной литературы.

СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обосновывается актуальность темы диссертационной работы, формулируется цель и решаемые задачи, выдвигаются основные защищаемые положения, определяется научная новизна результатов и их практическое значение.

В **первой главе** проведен анализ предметной области диссертационной работы, рассмотрены некоторые современные средства обучения иностранным языкам, их классификации, специфика и этапы развития. Приведена системная архитектура мультилингвистической адаптивно-обучающей технологии, проведен обзор ее информационно-алгоритмического обеспечения. Выявлены недостатки, основным из которых является неэффективное использование мощностных мультилингвистических словарей в процессе обучения.

Мощностной словарь в мультилингвистической адаптивно-обучающей технологии – это определенным образом организованная структура данных, содержащая в себе не только информацию о частотных характеристиках лексем, но и мощности связей между ними.

На основе таких словарей строятся средства информационной поддержки обучающих систем, так называемые информационно-терминологические базисы. На момент анализа предметной области мощности связей между понятиями языка использовались в модификации адаптивно-обучающего алгоритма для корректировки механизма формирования порций обучающей информации.

Критерий качества обучения МЛ-технологии имеет вид:

$$Q_n = \sum_{i=1}^{N} p_i^n(t_i^n)\xi_i \rightarrow \min_{p^n(t^n)}, \quad (1)$$

где $p_i^n(t_i^n)$ – вероятность незнания i-го элемента n-го набора ОИ;

t_i^n – время с момента последнего заучивания i-го элемента n-го набора ОИ;

ξ_i – абсолютная мощность лексической единицы в тексте, подвергшемуся частотной или обработке при составлении мощностного словаря, $0<\xi_i<1$;

$p^n(t^n)$ – вектор вероятностей незнания всех элементов n-го набора ОИ.

Для минимизации значения Q_n к концу сеанса обучения в порцию обучающей информации U_n^* включаются элементы базиса, имеющие наибольшее значение произведения $p^n(t^n)\ \xi_i$.

Если в оригинальном алгоритме мультилингвистической адаптивно-обучающей технологии роль абсолютной мощности лексической единицы играет абсолютная частота лексемы, то при нейросетевом подходе к формированию ИТБ абсолютная мощность лексической единицы определяется итерационно в процессе нейросетевой обработки текстов предметной области, с учетом лексических связей внутри изучаемого языка.

Такой подход акцентирует внимание обучаемого не только на лексемах с большей относительной частотой встречи, но и на так называемых центрах семантических сгущений.

Однако методика обучения языку остается прежней и представляет собой только последовательное запоминание терминов, в то время как само по себе наличие лексических связей дает возможность использовать сильнейшие механизмы обучения, а именно механизм установления однозначных ассоциативных зависимостей внутри изучаемого языка.

Таким образом, в первой главе обосновывается необходимость в разработке методики обучения многоязычной иностранной терминологии, ориентированной на формирование у обучаемого внутриязыковых ассоциативных связей непосредственно в процессе обучения.

Во **второй главе** вводится понятие лексически связанных компонентов (ЛС-компонентов), предлагается структура информационно-терминологического базиса как совокупности этих компонентов. На ее основе строится методика обучения с ориентацией на формирование ассоциативных связей между лексемами изучаемого языка (ЛСК-методика). Формулируются задачи о разработке алгоритмов формирования ИТБ как совокупности лексически связанных компонентов, и модификации адаптивного алгоритма обучения с учетом неоднородностей скоростей забывания терминологии.

Поскольку предлагаемая методика обучения должна акцентировать внимание не столько на изучении терминов, сколько на построении систем

ассоциативных связей внутри изучаемого языка(ов) возникает необходимость в пересмотре информационного базисного компонента ИТБ. В оригинальном случае он является совокупностью языковых аналогов и их частот (рисунок 1).

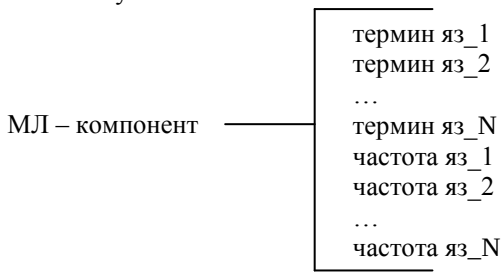
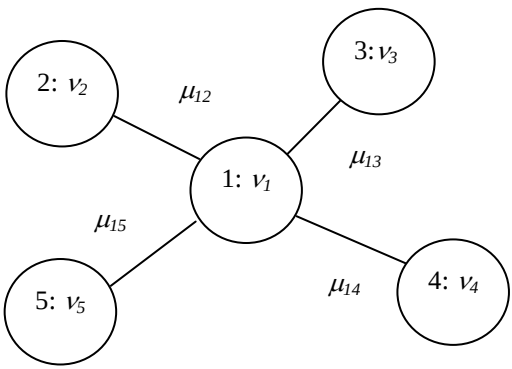


Рисунок 1 – Описание базисного информационного компонента с использованием нотации DSSD

В качестве информационного базисного компонента автор предлагает использовать объекты следующей структуры:



лексемы:
1 – основная лексема
2,3,4,5 – связанные лексемы
лексические связи:
1-2, 1-3, 1-4, 1-5
количественные характеристики:
ν_i – абсолютная частота *i*-ой лексемы;
μ_{ik} – абсолютная частота сочетания *i*–ой и *k*–ой лексем

Рисунок 2 – Лексически связанный компонент ИТБ

Такие комплексные объекты определены как лексически связанные компоненты (ЛС-компоненты) ИТБ.

Лексема, связанная со всеми без исключения лексемами ЛС-компонента называется основной лексемой, лексемы же, имеющие только одну связь – связанными лексемами.

Выражая структуру ЛС-компонента через МЛ-компонент получим:

ЛС-компонент = {МЛ-компонент (основная лексема), МЛ-компонент (связанная лексема №1), МЛ-компонент (связанная лексема №2),…} В нотации DSSD ЛС-компонент будет выглядеть:

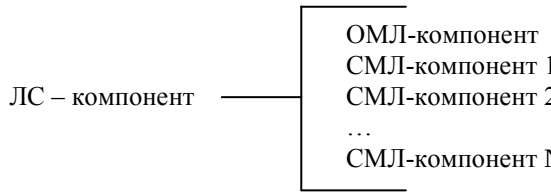


Рисунок 3 – Описание базисного информационного компонента с использованием нотации DSSD

При этом, структура МЛ-компонента, базирующегося на основной лексеме (основного) должна быть откорректирована, иначе теряется информация о лексических связях, а значит и смысл всей методики обучения.

Структура основного МЛ-компонента (ОМЛ-компонента) будет выглядеть следующим образом: ОМЛ-компонент = {термин яз_1, термин яз_2,..., термин яз_N, частота яз_1, частота яз_2,..., частота яз_N, сочетание 1_яз_1, сочетание 1_яз_2,..., сочетание 2_яз_1, сочетание 2_яз_2,…, сочетание K_яз_N, частота 1_яз_1, частота 1_яз_2,…, частота 2_яз_1, частота 2_яз_2,…, частота K_яз_N}.

Задача формирования ЛС-компонентов и ИТБ на их основе имеет два аспекта, которые должны быть учтены при разработке соответствующих алгоритмов:

– должен быть выработан критерий, согласно которому из терминологической базы будут выделятся основные лексемы;

– количество связей в ЛС-компоненте должно быть оптимальным.

Методика обучения иностранной лексике строится на совместном применении двух алгоритмов: адаптивного алгоритма обучения и алгоритма построения внутриязыковых ассоциативных полей.

Алгоритм построения внутриязыковых ассоциативных полей состоит в последовательной подаче к изучению элементов ЛС-компонента. Последовательность такова:

1. Основная лексема – перевод, подсказка на иностранном языке.
2. Связанная лексема – перевод, подсказка на иностранном языке.
3. Лексическое сочетание основной и связанной лексем – перевод сочетания, подсказка на иностранном языке (языковой аналог именно лексического сочетания, но не лексем по отдельности).
4. Переход к следующей связанной лексеме.
5. Переход к следующему лексически связанному компоненту.

При использовании данной методики адаптивный алгоритм обучения должен быть модифицирован. Во-первых, основным элементом обучающей информации будет не термин, но лексически связанный компонент; во-вторых, согласно предложенной автором методики, скорость забывания для основных и связанных лексем будет не одинакова. Основные лексемы ЛС-компонентов

будут заучены гораздо лучше, чем связанные. Это обуславливается ассоциативными механизмами памяти человека, а именно тем, что при актуализации одного понятия актуализируется другое, ассоциативно связанное с ним; вместе с этим предлагаемая автором методика такова, что наибольшее количество ассоциативных связей приходится именно на основные лексемы. Простое повторение терминов не задействует ассоциативные механизмы памяти и не будет столь эффективным, в то время как обучение по данной методике формирует строго организованные системы ассоциативных связей с наиболее значимыми центрами.

В **третьей главе** предлагаются алгоритмы формирования информационно-терминологического базиса как совокупности лексически связанных компонентов, описана модификация адаптивного алгоритма

Разработка алгоритмов формирования ИТБ в конечном итоге имеет своей целью построение таких ИТБ, использование которых в процессе обучения будет высоко эффективным. Поэтому при разработке данных алгоритмов рационально отталкиваться непосредственно от процесса обучения, который в достаточной мере описывается адаптивным алгоритмом. Следовательно, формирование алгоритмической поддержки предложенной методики разумно начать с его модификации, проведя корректировку для итерационного вычисления скоростей забывания в процессе адаптации параметров модели к ответам обучаемого, а так же изменив способ формирования порций обучающей информации.

Формула для расчета скоростей забывания в предлагаемой модификации будет выглядеть следующим образом:

$$\alpha_i^n = \frac{b_i^n}{(\hat{h}k\frac{\sum_{j=1}^{k}(1-p_{ij})\mu_{ij}}{\sum_{j=1}^{k}\mu_{ij}}+1)}, \qquad (2)$$

где b_i^n - скорость забывания i-го элемента n-го набора ОИ;

$\hat{h}$ – оценка параметра h, характеризующего активность ассоциативных связей,

$0 < \hat{h} < 1$;

k – количество связанных лексем в компоненте;

$(1-p_{ij})$– вероятность знания j-ого элемента ОИ, который лексически связан (т.е. порождает ассоциацию), с i-ым элементом ОИ;

μ_{ik} – абсолютная частота сочетания i–ой и k–ой лексем.

Отметим, что $\frac{\sum_{j=1}^{k}(1-p_{ij})\mu_{ij}}{\sum_{j=1}^{k}\mu_{ij}}$ – есть не что иное, как средневзвешенная вероятность знания связанных лексем компонента.

Параметр h характеризует вероятность актуализации основной лексемы при актуализации связанной с ней согласно структуре ЛС-компонента.

Данный параметр зависит не столько от изучаемого языка и организации информационно терминологического базиса, сколько от самого обучаемого и ситуации, в которой происходит актуализация связанных лексем. Если речь идет о чтении книги на языке, которым читатель владеет свободно, то параметр h близок к 0, поскольку нет необходимости в порождении ассоциаций при прочтении каждого слова. Но если речь идет о переводе текста, особенно в том случае, когда переводчик слабо знаком с терминологией, значение параметра h резко возрастает, поскольку объектом внимания переводчика являются не сложные лексические конструкции, такие как предложения и текст в целом, но сами термины и их смысл. В нашем случае параметр h близок единице. Так или иначе, речь идет только о процессе обучения, это означает, что h будет зависеть только от личности обучаемого. Поэтому на каждой итерации, при вычислении скоростей забывания вместо h разумно брать ее простейшую оценку по ответам обучаемого, а именно:

$$\hat{h} = \frac{m_h}{n_h}, \tag{3}$$

где m_h – количество запомненных словосочетаний при знании соответствующих связанных лексем;

n_h – общее количество словосочетаний, выданных для проверки, при знании соответствующих связанных лексем.

Что касается формирования порций обучающей информации, то в целях получения устойчивых групп ассоциативно связанных элементов к концу обучения, были сформулированы следующие корректировки к оригинальному алгоритму:

– если в обучающую порцию попадает связанная лексема, то для повторного изучения она выдается в связке с основной;

– если в обучающую порцию попадает основная лексема, то для повторного изучения выдается весь ЛС-компонент.

Примечательно, что при использовании полученных знаний, в зависимости от параметра h, при встрече связанной лексемы, обучаемый будет вспоминать основную, что резко сократит вероятность забывания основных лексем по сравнению с оригинальным мультилингвистическим подходом.

Далее в третьей главе рассматривается нисходящий алгоритм формирования ЛС-компонентов, учитывающий как частоту терминов, так и мощность их лексических связей. Данный алгоритм можно разделить на три основные фазы.

1. Подготовка ИТБ.

1.1 Для каждой лексемы ИТБ вычисляется значение коэффициента значимости по следующей формуле:

$$L_i = \mathrm{e}^{-\frac{0,7}{k\frac{\sum_{j=1}^{k} q_j \mu_{ij}}{\sum_{j=1}^{k} \mu_{ij}}+1}} q_i, \qquad (4)$$

где μ_{ik} – относительная частота сочетания *i*–ой и *k*–ой лексем, отражает силу ассоциативной связи;

q_i – относительная частота, выражающая долю лексической единицы в тексте, подвергшемуся статистической обработке при составлении частотного словаря, $0<q_i<1$.

1.2 ИТБ упорядочивается по убыванию значения L_i (таким образом, чем меньше будет порядковый номер лексемы, тем выше вероятность, образования на ее основе ЛС-компонента).

1.3 Данные о лексических связях упорядочиваются по убыванию значения $q_k \mu_{ik}$ (тем самым увеличивается вероятность попадания в ЛС-компонент тех из связанных лексем, которые более всего могут улучшить качество ИТБ).

2. Поиск оптимального количества основных лексем.

2.1 Осуществляется перебор возможного количества основных лексем k от 1 до объема ИТБ (возможно сужение интервала поиска разработчиком).

2.2 Для текущего значения k, определяются основные лексемы (k первых лексем ИТБ).

2.3 Для выбранных основных лексем определяются связанные лексемы (как правило, задается максимум количества связанных лексем).

2.4 Вычисляется значение функции качества ИТБ согласно формуле:

$$L(n) = \sum_{i=1}^{N} q_i \mathrm{e}^{-\frac{0,7}{k\frac{\sum_{j=1}^{k} q_j \mu_{ij}}{\sum_{j=1}^{k} \mu_{ij}}+1}}. \qquad (5)$$

L(n) показывает сумму взвешенных вероятностей знания лексем по всему базису, естественно, что чем больше эта сумма, тем более удачно построен базис.

2.5 Если перебор окончен, идем в пункт 2.6, иначе возврат к пункту 2.1.

2.6 Определяем максимум функции качества от числа ЛС-компонентов (оптимальное число основных лексем k_{max}).

3. Формирование ИТБ как совокупности ЛС-компонентов (искомый ИТБ получается при прохождении пунктов 2.2 и 2.3 для k_{max} основных лексем).

Нисходящий алгоритм формирования ЛС-компонентов дает «хорошие» результаты, однако из пунктов 2.2 и 2.3 следует, что связанные лексемы определяются согласно порядка основных лексем. Естественно, что связанная

лексема, являясь частью одного ЛС-компонента, уже не может быть частью другого, даже если она "подходит" к нему больше (речь идет о конечном значении *L(n)*). Таким образом, возникает задача о нахождении наиболее "подходящих" связанных лексем для ЛС-компонентов в процессе их формирования. Эта задача решается от обратного, т.е. не подбирая для основных лексем связанные, но наоборот. Алгоритм, работающий по такому принципу получил название восходящего алгоритма формирования ЛС-компонентов и так же может быть поделен на три основные фазы:

1. Подготовка ИТБ.

1.1 Для каждой лексемы ИТБ вычисляется значение L_i.

1.2 ИТБ упорядочивается по убыванию значения L_i.

2. Поиск оптимального количества основных лексем.

2.1 Осуществляется перебор возможного количества основных лексем k от 1 до объема ИТБ (возможно сужение интервала поиска разработчиком).

2.2 Для текущего значения k, определяются основные лексемы (k первых лексем ИТБ).

2.3 Осуществляется перебор "не основных" (потенциально связанных) лексем и для каждой "не основной" лексемы l рассчитывается множество значений приращения функции качества, полученные путем ее включения во все возможные ЛС-компоненты. Расчет осуществляется по формуле:

$$\Delta L_i = q_i \mathrm{e}^{-\frac{0,7}{(k+1)\frac{q_l\mu_{il}+\sum_{j=1}^{k} q_j\mu_{ij}}{\mu_{il}+\sum_{j=1}^{k}\mu_{ij}}+1}} - L_i^k, \qquad (6)$$

где L_i^k – коэффициент значимости i-ой основной лексемы, образующей ЛС-компонент с k связанными лексемами;

Далее определяется ЛС-компонент для которого ΔL_i будет максимальным, лексема l включается в его состав в качестве связанной (как правило, задается максимум количества связанных лексем).

2.4 Подсчитывается значение функции качества.

2.5 Если перебор окончен, идем в пункт 2.6, иначе возврат к пункту 2.2.

2.6 Определяем максимум функции качества от числа ЛС-компонентов (оптимальное число основных лексем k_{max}).

3. Формирование ИТБ как совокупности ЛС-компонентов.

3.1 Незадействованные в ЛС-компонентах лексемы из числа основных (k_{max}) помечаем как "не основные". Нахождение наиболее "подходящих" связанных лексем порождает свободные элементы из числа потенциально основных лексем, что во многом ухудшает *L(n)*; поэтому незадействованные в ЛС-компонентах лексемы из числа основных (k_{max}) помечаем как "не основные".

3.2 Для полученного значения k_{max} основных лексем осуществляем шаги 2.2 и 2.3 и получаем тем самым искомый ИТБ.

Так же в третьей главе приводится сравнительный анализ предлагаемых алгоритмов на 3-х ИТБ одинаковой структуры, но различного объема.

Настраиваемые параметры базиса:

- максимальное количество связей, приходящихся на одну лексему (10);
- максимальное значение абсолютной частоты лексем (100/50000);
- максимальное значение частоты сочетаний лексем (20/50000);
- объем материала, по которому произведен частотный анализ(50000);
- коэффициент связанности лексем (1);

Тест 1 (объем базиса 1000 терминов)

Параметры выхода	Н-алгоритм	В-алгоритм
min L(n)	0,4961185136	0,49611851363
max L(n)	0,4961185651	0,49611856954
Основных лексем	188	172
Свободных элементов	320	324
Итераций	4981096	2000498500
Время исполнения (сек)	3,41	130,75

Таблица 1 – Результаты теста 1

Тест 2 (объем базиса 2000 терминов)

Параметры выхода	Н-алгоритм	В-алгоритм
min L(n)	0,99479940765	0,99479940765
max L(n)	0,99479951781	0,99479952460
Основных лексем	410	394
Свободных элементов	545	538
Итераций	19930343	21542653063
Время исполнения (ч:м:с:сс)	00:00:18:22	00:23:28:16

Таблица 2 – Результаты теста 2

Тест 3 (объем базиса 5000 терминов)

Параметры выхода	Н-алгоритм	В-алгоритм
min L(n)	2,4771264022	2,4771264022
max L(n)	2,4771266487	2,4771266614
Основных лексем	845	822
Свободных элементов	1867	1886
Итераций	124568303	253340376005
Время исполнения (ч:м:с:сс)	00:01:17:50	04:35:58:57

Таблица 3 – Результаты теста 3

Приведем сводную таблицу результатов тестирования для времени исполнения и *L(n)*.

Показатели качества	Н-алгоритм	В-алгоритм
$\Delta L(n)$ (1000)	0,00000005150	0,00000005590
$\Delta L(n)$ (1000), %	100	108,54
$\Delta L(n)$ (2000)	0,00000011016	0,00000011695
$\Delta L(n)$ (2000), %	100	106,16
$\Delta L(n)$ (5000)	0,00000024658	0,00000025921
$\Delta L(n)$ (5000), %	100	105,12
Δt (1000), %	100	$38{,}34 * 10^2$
Δt (2000), %	100	$76{,}7 * 10^2$
Δt (5000), %	100	$213{,}65 * 10^2$

Таблица 4 – Сводная таблица результатов тестирования

В таблице 4 приведены результаты 3-х экспериментов ($\Delta L(n) = max\ L(n) - min\ L(n)$), согласно которым В-алгоритм превосходит Н-алгоритм на 8,54; 6,16; 5,12 %, соответственно. Снижение этого превосходства обусловлено негативным влиянием свободных элементов из числа основных лексем на *L(n)*, которое становится сильнее с ростом ИТБ и постепенно подавляет положительное влияние "наиболее подходящих" связанных лексем. Поскольку ИТБ, с которыми работает ЛСК-методика, предметно-ориентированы и их объем, как правило, не превышает 5000 терминов, ограничимся на том, что В-алгоритм превосходит Н-алгоритм по качественным показателям на 5-10 %.

Что касается времени исполнения, то здесь заметна тенденция: при увеличении объема ИТБ в n раз, время исполнения алгоритмов возрастает в геометрической прогрессии, причем для В-алгоритма это время возрастает в *n* раз быстрее (исключая помехи на больших объемах ИТБ).

Несмотря на большую, негативную разницу по времени исполнения, В-алгоритм формирует более качественную структуру ИТБ. Поскольку операция формирования ИТБ производится только один раз, показатель качества много важнее времени исполнения алгоритма.

В-алгоритм может успешно использоваться при формировании ИТБ как совокупности ЛС-компонентов. Единственным его серьезным недостатком является не время исполнения, а то, что заранее невозможно предугадать, сколько именно основных лексем будет в ИТБ. В случае, когда разработчик выставляет жесткие требования к количеству основных лексем (ЛС-компонентов) или имеется значительный объем ИТБ, следует использовать Н-алгоритмы формирования ЛС-компонентов, в противном случае, предпочтение следует отдавать В-алгоритмам.

В **четвертой главе** описывается структура и основные аспекты формирования ИТБ в виде электронного двухблочного лексически связанного словаря как практическое применение средств алгоритмической поддержки предлагаемой методики. Так же обозначены основные направления применения на практике структур данных, алгоритмов и методики обучения разработанных в ходе написания диссертационной работы.

Двухблочная структура предлагаемого в четвертой главе словаря (особенно в печатном варианте) делает его более универсальным, сохраняя его старый функционал и расширяя возможности обучения за счет применения ЛСК-методики.

Первый блок, изображенный на рисунке 4, содержит лексически связанные компоненты, которые представлены в виде двухуровневых структур данных, где на верхнем уровне находятся основные лексемы (крупный шрифт), на нижнем – связанные и сочетания (мелкий шрифт).

		A		
2	1	accuracy, 47	Genauigkeit, 35 f	точность, 139
15	14	action, 164	Handlung, 90 f	действие, 172
22	21	activity, 10	Aktivität, 4 f	деятельность, 102
25	24	adder, 51	Addierer, 47 m	сумматор, 127
38	37	address, 168	Adresse, 130 f	адрес, 168
51	50	algebra, 86	Algebra, 43 f	алгебра, 34
60	59	algorithm, 169	Algorithmus, 89 m	алгоритм, 126
73	72	analysis, 197	Analyse, 169 f	анализ, 327
74	73	comprehensive, 30	umfassend, 9	всесторонний, 30
75	74	comprehensive analysis, 9	umfassende Analyse, 7 f	комплексный анализ, 23
76	75	expert, 23	Experte, 3 m	эксперт, 55
77	76	expert analysis, 11	Expertenanalyse, 9 f	экспертный анализ, 26
78	77	factor, 68	Faktor, 44 m	коэффициент, 47, фактор, 31
79	78	factor analysis, 12	Faktoranalyse, 10 f	факторный анализ, 4
80	79	job, 50	Arbeit, 25 f	работа, 179
81	80	job analysis, 3	Arbeitsanalyse, 9 f	анализ работы, 13
82	81	network, 36	Netzwerk, 13 n	сеть, 9
83	82	network analysis, 9	Netzwerkanalyse, 2 f	сетевой анализ, 14
84	83	qualitative, 2	qualitativ, 6	качественный, 33
85	84	qualitative analysis, 8	qualitative Analyse, 8 f	качественный анализ, 19
86	85	approach, 70	Herangehensweise, 8 f, Methode, 24 f	подход, 60

Рисунок 4 – Блок лексически связанных компонентов

Второй блок словаря, изображенный на рисунке 5, содержит всю терминологию, включая лексемы, вошедшие в лексически связанные компоненты; они выделены курсивом и формально представляют собой ссылки на соответствующие элементы первого блока (поиск осуществляется по нумерации).

2772	2186	vulgar fraction, 2	gewöhnlicher Bruch, 2 m	простая дробь, 2
2773	2187	vulgar, 2	vulgär, 2	вульгарный, 2
2774		***W***		
2775	***540***	***waiting, 8***	***Warten, 8 n***	***ожидание, 8***
2776	***541***	***waiting time, 8***	***Wartezeit, 2 f***	***время ожидания, 6***
2777	2188	wave function, 2	Wellenfunktion, 2 f	волновая функция, 2
2778	2189	wave, 12	Welle, 7 f	волна, 9
2779	2190	welfare, 3	Wohl, 6 n	благополучие, 2
2780	2191	white noise, 8	weißes Rauschen, 2 n	белый шум, 3
2781	2192	white, 8	weiß, 2	белый, 3
2782	2193	whole, 15	ganz, 9	целое, 69
2783	2194	wholeness, 6	Ganzheit, 2 f	целостность, 23
2784	***48***	***word, 22***	***Wort, 3 n***	***слово, 22***
2785	***49***	***word address, 22***	***Wortadresse, 3 f***	***адрес слова, 2***
2786	2195	work file, 3	Arbeitsdatei, 3 f	рабочий файл, 2
2787	2196	work, 3	Arbeit, 24	работа, 24
2788	2197	working space, 3	Arbeitsraum des Speichers, 2 m	рабочая область памяти, 33
2789	***127***	***working, 9***	***Arbeiten, 9 n***	***работа, 24***
2790	***128***	***working cell, 3***	***Arbeitszelle, 6 f***	***рабочая ячейка, 4***
2791	2198	world, 7	Welt, 3 f	мир, 111
2792	2199	write operation, 4	Schreiboperation, 3 f	операция записи, 4
2793	***542***	***write, 12***	***schreiben, 12***	***написать, 12***
2794	***543***	***write time, 12***	***Schreibzeit, 4 f***	***время записи, 9***
2795		***X***		
2796	2200	X-axis, 4	Abszissenachse, 2 f	ось абсцисс, 2
2797		***Y***		
2798	2201	Y-axis, 4	Ordinatenachse, 2 f	ось ординат, 2
2799		***Z***		
2800	***275***	***zero, 65***	***Null, 4 f***	***нуль, 2***
2801	***276***	***zero level, 15***	***Nullniveau, 10 n***	***нулевой уровень, 14***
2802	2202	zonal, 15	zonal, 1	зональный, 1
2803	2203	zone, 41	Zone, 4 f	зона, 7
2804	2204	zoom, 12	zoomen, 20 n	масштабировать, 6

Рисунок 5 – Терминологический блок лексически связанного словаря

Терминологический блок используется в качестве справочного материала для перевода текстов иностранных терминологий, так же рекомендуется для обучения, согласно мультилингвистического подхода.

Оба блока упорядочены по алфавиту, добавлены литеры. Группировка осуществлена средствами excel; для первого блока является трехуровневой (литера – основная лексема – связанные лексемы), для второго – двухуровневой (литера – основная лексема).

По отношению к своему предшественнику, англо-немецко-русскому частотному словарю по информатике и системному анализу, предлагаемый словарь был дополнен многоязычной терминологией в количестве 300 специальных терминов и адаптирован для обучения по средством методики построения внутриязыковых ассоциативных полей.

Несмотря на то, что ЛС-компоненты были разработаны специально для информационно-терминологической поддержки ЛСК-методики, они могут эффективно применяться в построении информационных систем и структур данных, которые эту методику не поддерживают, примером этого могут служить мультилингвистические поисковые системы, что во многом определяет дальнейшие перспективы развития диссертационного исследования.

Таким образом, четвертая глава, раскрывает практическое применение разработанных автором алгоритмов и структур данных, определяет возможные перспективы развития данного исследования, и является логическим завершением диссертационной работы в целом.

В **заключении** сформулированы основные выводы и результаты, полученные в диссертационной работе.

Основные научные результаты и выводы.

1. Проведен теоретико-информационный анализ методов и алгоритмов мультилингвистической адаптивно-обучающей технологии как сложной проблемно-ориентированной системы управления, в ходе которого выявлена проблема построения внутриязыковых ассоциативных полей непосредственно в процессе обучения. Рассмотрены возможные пути ее решения.

2. Проведен теоретико-информационный анализ структуры информационной базы мультилингвистической адаптивно-обучающей технологии, что обеспечило возможность ее последующей реорганизации и применения на ее основе разработанной в ходе диссертационного работы методики обучения на базе лексически связанных информационных компонентов.

3. Как совокупность алгоритмов и методов управления сложным обучающим процессом разработана методика обучения многоязычной иностранной терминологии, позволяющая непосредственно в ходе обучения формировать у объекта строго организованные системы внутриязыковых ассоциативных связей на всем множестве языков, задействованных в обучении. Применение методики на практике делает конечное знание более структурированным, облегчает процесс его интеграции в жизнь, снижает скорость забывания терминологии в целом.

4. Разработана структура частотных мультилингвистических лексически связанных словарей как средства информационного обеспечения предложенной методики, что делает возможным ее применение в учебном процессе, организованном самостоятельно или с участием преподавателя.

5. Разработана модификация адаптивного алгоритма обучения на основе лексически связанных словарей, формирующая в процессе своей работы строго организованные системы внутриязыковых ассоциативных связей, учитывающая неоднородность скоростей забывания. Разработка данной модификации обеспечивает возможность успешной интеграции предлагаемой методики обучения в интерактивные мультилингвистические адаптивно-обучающие компьютерные системы.

6. Разработан и программно реализован комплекс алгоритмической поддержки предложенной методики, в том числе алгоритмы структурно-параметрического синтеза информационно-терминологического базиса как совокупности лексически связанных компонентов. Данный программно-алгоритмический комплекс нашел свое применение при создании и последующей обработке двухблочного мультилингвистического лексически связанного словаря как средства информационной поддержки предлагаемой методики обучения.

7. Согласно предложенной методике обучения на основе разработанных алгоритмов и других компьютерных методов обработки информации сформирован информационно-терминологический базис в форме двухблочного мультилингвистического лексически связанного словаря по

информатике и системному анализу. Словарь реализован в электронном и печатном виде, откорректирован и расширен по сравнению с предыдущей версией. Специально разработанная двухблочная структура словаря позволяет эффективно применять его как в «классическом» процессе обучения, так и в обучении посредством разработанной автором методики.

Основные результаты диссертационной работы **опубликованы** в следующих работах:

1. * Лесков, В.О. Алгоритмизация процедур включения связанных лексем в структуру информационно-терминологического базиса / М.В. Карасева, И.В. Ковалев, В.О. Лесков // Программные продукты и системы. – 2009. – №3. – С. 35–38.
2. * Лесков, В.О. Компоненты информационной поддержки мультилингвистической адаптивно-обучающей технологии / М.В. Карасева, И.В. Ковалев, В.О. Лесков // Системы управления и информационные технологии. – 2009. – №1.3 (35). – С. 360–363.
3. * Лесков, В.О. Процедура построения частотного словаря на основе лексически связанных компонентов / И.В. Ковалев, В.О. Лесков, Е.Е. Шукшина // Вестник СибГАУ. – 2009. – №2(23). – С. 119–122.
4. * Лесков, В.О. Формирование лексически связанных компонентов информационно-терминологического базиса / В.О. Лесков // Вестник СибГАУ. – 2009. – №2(23). – С. 133–136.
5. * Лесков, В.О. Информационно-терминологический базис как совокупность лексически связанных компонентов / М.В. Карасева, И.В. Ковалев, В.О. Лесков // Вестник СибГАУ. – 2009. – №1(22). – С. 54–56.
6. * Лесков, В.О. Адаптивный алгоритм обучения иностранной лексике на основе лексически связанных компонентов / М.В. Карасева, И.В. Ковалев, В.О. Лесков // Системы управления и информационные технологии. – 2009. – №4(34). – С. 78–82.
7. * Лесков, В.О. Внутриязыковые ассоциативные поля в мультилингвистической адаптивно-обучающей технологии / М.В. Карасева, И.В. Ковалев, В.О. Лесков // Системы управления и информационные технологии. – 2008. – №3.1(33). – С. 157–160.
8. * Лесков, В.О. Автоматизация формирования информационной базы мультилингвистической адаптивно-обучающей технологии / М.В. Карасева, В.О. Лесков // Вестник СибГАУ. – 2007. – №4(17). – С. 117–124.
9. Лесков, В.О. Специфика модели обучения на основе лексически связанных компонентов /В.О. Лесков // Современные наукоёмкие технологии. – 2009. – №4. – С. 53–54.
10. Лесков, В.О. Два блока частотного словаря: значение и организация / В.О. Лесков // Успехи современного естествознания. – 2009. – №4. – С. 31–32.

---

* работа, опубликованная в одном из ведущих рецензируемых научных журналов, рекомендуемых ВАК РФ для опубликования основных научных результатов диссертационных исследований

11. Лесков, В.О. Область применения лексически связанных компонентов /В.О. Лесков // Современные наукоёмкие технологии. – 2009. – №2. – С. 63–64
12. Лесков, В.О. Два подхода к формированию лексически связанных компонентов /В.О. Лесков // Современные наукоёмкие технологии. – 2008. – №12. – С. 29–30.
13. Лесков, В.О. О путях искуственного формирования сложных ассоциативных связей в процессе обучения иностранной лексике. /В.О. Лесков //Современные наукоёмкие технологии. – 2008. – №4. – С. 78–79.
14. Лесков, В.О. Реляционная модель и алгоритмы оптимизации модульной структуры мультилингвистического информационно-терминологического базиса /В.О. Лесков, С.С. Огнерубов// Вестник Университетского Комплекса. – 2006. – №7(21). – С. 116–133.
15. Лесков, В.О. Информационная модель динамического формирования электронных терминологических словарей / К.А. Дудура, В.О. Лесков, С.С. Огнерубов // Недра Кузбасса. Инновации: Труды V Всероссийской научно-практической конференции. – Кемерово: ИНТ, 2006. – С. 75–76.

Разработки, зарегистрированные в Отраслевом фонде алгоритмов и программ:

1. Лесков, В.О. Адаптивно-обучающий алгоритм ЛСК-методики / В.О. Лесков. – М.:ВНТИЦ, 2009. – № 50200900256.
2. Лесков, В.О. Комплекс программного моделирования КПМ v.2.0 / В.О. Лесков. – М.:ВНТИЦ, 2009. – № 50200900124.
3. Лесков, В.О. Двухблочный трехуровневый элекстронный англо-немецко-русский частотный словарь по информатике и системному анализу / М.В. Карасева, И.В. Ковалев, В.О. Лесков. – М.:ВНТИЦ, 2009. – № 50200900111.
4. Лесков, В.О. Нисходящий алгоритм формирования ЛС-компонентов / В.О. Лесков. – М.:ВНТИЦ, 2008. – № 50200802428.
5. Лесков, В.О. Восходящий алгоритм формирования ЛС-компонентов / В.О. Лесков. – М.:ВНТИЦ, 2008. – № 50200802427.
6. Лесков, В.О. Комплекс программного моделирования КПМ v.1.0 / В.О. Лесков. – М.:ВНТИЦ, 2008. – № 50200802242.
7. Лесков, В.О. Программа анализа и формирования информационного мультилингвистического терминологического базиса, на основе реляционной модели оптимизации TuMLas v.1.0 / И.В. Ковалев, В.О. Лесков. – М.:ВНТИЦ, 2008. – № 50200701283.
8. Лесков, В.О. Программа контекстного анализа методом «Скрытых Марковских цепей» / И.В. Ковалев, В.О. Лесков. – М.:ВНТИЦ, 2008. – № 50200501669.

Лесков Виталий Олегович

Мультилингвистические системы адаптивного обучения на базе лексически связанных информационных компонентов

Автореф. дисс. … канд. техн. наук

Подписано в печать 08.10.09. Тираж 100 экз.

Отпечатано в типография КрасГМУ
660022, г. Красноярск, ул. П.Железняка, 1