УДК 004.912

С.Ф. Липницкий, А.А. Мамчич, Л.В. Степура
Объединенный институт проблем информатики
Национальной академии наук Беларуси

ЭЛЕКТРОННЫЙ БАНК ДАННЫХ ДЛЯ ИНФОРМАЦИОННОГО ОБЕСПЕЧЕНИЯ СПЕЦИАЛИСТОВ В ОБЛАСТИ НАНОТЕХНОЛОГИЙ

В работе представлены результаты проекта по созданию электронного банка данных в области нанотехнологий. Предлагается архитектура банка данных и рассмотрены принципы функционирования его программных компонентов, обеспечивающих поиск и обработку текстовых документов из различных информационных источников.

У роботі представлені результати проекту по створенню електронного банку даних в галузі нано-технологій. Пропонується архі-тектура банку даних і розглянуті принципи функціонування його програмних компонентів, що забезпечують пошук і обробку текстових документів з різних інформаційних джерел.

The results of project are in-process presented on creation of electronic bank of the нанотехнологий given in area of. Architecture of bank of data is offered and principles are considered of functioning of its software components, providing a search and treatment of text documents from different informative sources.

Ключевые слова: электронный банк данных, архитектура электронного банка данных, индексирование текстовых документов

Введение

В настоящее время в связи с интенсивным увеличением объемов текстовой информации, представленной в электронном виде, все более необходимой становится разработка систем, направленных на обеспечение информационных потребностей специалистов в конкретных предметных областях. Электронный банк данных представляет собой систему централизованного хранения и коллективного использования актуальных знаний в области нанотехнологий. В состав банка данных входит комплекс информационных, технических, программных, языковых и организационных средств, обеспечивающих сбор, хранение, поиск и обработку текстовых документов из различных источников (сеть Интернет, локальная сеть, жесткий диск отдельных компьютеров). В отличие от существующих программных систем, в которых используются технологии, ориентированные на исследование структуры и статистических характеристик самих документов без привлечения дополнительной информации [1-3], в электронном банке данных в качестве знаний о предметной области применяются тематические корпусы текстов и сформированные на их основе лингвистические словари. Такая методика обеспечивает адаптацию системы к решаемой задаче и независимость программного комплекса от входных языков. Данные корпусы текстов могут создаваться предварительно, а также формироваться в оперативном режиме непосредственно при поиске информации путем объединения наборов документов, релевантных каждому конкретному тексту или запросу пользователя (так называемые динамические корпусы).

1. Архитектура электронного банка данных по нанотехнологиям

Функциональными компонентами системы электронного банка данных по нанотехнологиям являются три подсистемы (рис. 1):

– автоматизированное рабочее место (АРМ) оператора банка данных;

– подсистема поиска текстовых документов в различных информационных источниках;

– подсистема реферирования найденных документов.

1.1. Автоматизированное рабочее место оператора банка данных

АРМ оператора банка данных – это программно-информационный инструментарий, предназначенный для визуализации процедур создания и ведения баз данных и знаний и управления этими процедурами. В базе данных оператор накапливает и классифицирует по различным тематическим разделам предметной области нанотехнологий электронные варианты документов для последующего создания на их основе тематических корпусов. (Тематический корпус текстов – это совокупность полнотекстовых документов, посвященных какой-либо конкретной тематике.) На основе информации из базы данных в системе формируются лингвистические словари, используемые при реализации функций поиска и реферирования документов из различных информационных источников.

1.2. Подсистема поиска текстовых документов

Основными функциями подсистемы являются индексирование текстовых документов из различных информационных источников на основе использования словарей базы знаний и осуществление процедуры поиска по запросам пользователей. Основными компонентами подсистемы являются:

– программа индексирования текстовых документов. Программа приписывает каждому документу совокупность ключевых слов (дескрипторов) и их весовых коэффициентов на основе статистического анализа частот этих слов в индексируемом документе и полном корпусе текстов, образованном всеми тематическими корпусами;

– база данных – хранилище поисковых образов загруженных и проиндексированных документов;

– программа выдачи результатов индексирования из базы данных.

1.3. Подсистема реферирования текстовых документов

В процессе реферирования происходит выявление наиболее информативных предложений текста с целью краткого ознакомления с ним пользователя электронного банка данных по нанотехнологиям. В состав подсистемы входят следующие основные структурные компоненты:

– база знаний, включающая систему лингвистических словарей и совокупность тематических корпусов текстов;

– программа поиска информативных предложений, которая выявляет в тексте информативные предложения. Количество таких предложений регулируется путем задания пользователем порогового уровня информативности. При этом возможно предъявление пользователю (по его требованию) контекста каждого информативного предложения.

Результатом работы подсистемы реферирования является кортеж информативных предложений упорядоченных в порядке их расположения в соответствующем документе.

2. Индексирование текстовых документов

Процедура индексирования текстовых документов реализуется в два этапа [4]. На первом этапе происходит предварительная обработка документа *t*, которая включает в себя лексический анализ текста (удаление элементов форматирования, цифр, элементов пунктуации, математических формул и т. п.), исключение стоп-слов (слов малой информативности, которые не нужно учитывать при поиске документов: союзы, местоимения и т. д.). На втором этапе формируется поисковый образ в виде множества ключевых слов документа с определенными числовыми коэффициентами (обычно их называют весами ключевых слов, которые определяют их информативность).

2.1. Тематические и динамические корпусы текстов

Пусть имеется некоторое непустое множество текстов (совокупность текстов по конкретной тематике). Сформируем текст *Ct*, объединив все множества предложений каждого из этих текстов, и назовем его тематическим корпусом текстов**.**

Поскольку в информационной системе представлено, как правило, несколько таких корпусов, будем обозначать их *Cti* (*i* – номер корпуса). Объединение $Cf = \bigcup_{i=1}^{n} Ct_i$ всех тематических корпусов назовем полным корпусом текстов.

В процессе поиска текстовых документов с целью адаптации к информационным потребностям пользователей осуществляется коррекция их первоначальных запросов. Данная процедура реализуется с использованием динамических корпусов текстов. Под динамическим будем понимать корпус текстов первого порядка, все документы которого релевантны некоторому текстовому документу или запросу на поиск информации.

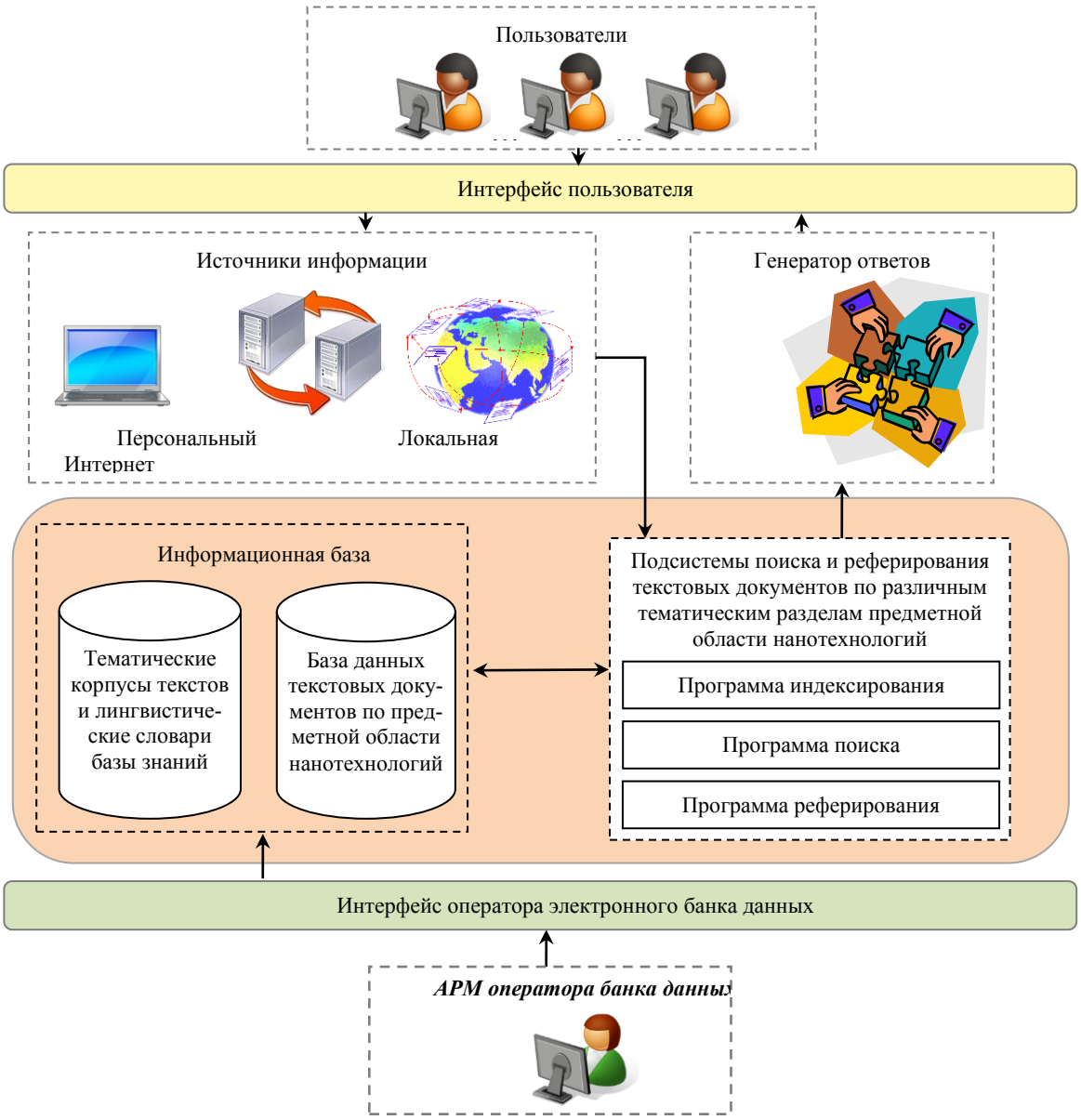

Рис. 1. Архитектура электронного банка данных о нанотехнологиях

2.2. Вычисление информативности словоформ

Информативность словоформ вычисляется с использованием результатов статистической обработки тематических или динамических корпусов текстов и полного корпуса текстов [5]. Рассмотрим процесс вычисления информативности словоформ на основе тематических корпусов текстов с учетом того, что для динамических корпусов он осуществляется аналогичным образом.

Рассмотрим следующую совокупность событий:

SCt – словоформа извлечена случайным образом из тематического корпуса текстов *Ct* *(Ct* $\in$ *Cf)*;

SCf – словоформа извлечена из полного корпуса текстов *Cf*;

HCt – появление тематического корпуса текстов *Ct*.

Пусть *P(SCt* / *SCf)* – условная вероятность того, что словоформа α извлечена из множества *Ct* при условии, что он уже извлечена из полного корпуса текстов *Cf*. Эта вероятность вычисляется следующим образом:

$$P(S_{Ct} / S_{Cf}) = \frac{P(S_{Ct} \cdot S_{Cf})}{P(S_{Cf})} = \frac{P(S_{Ct}) \cdot P(S_{Cf} / S_{Ct})}{P(S_{Cf})}.$$

Вероятность $P(SCt/SCf)$ будем называть информативностью словоформы α в тематическом корпусе текстов Ct.

При достаточно больших объемах полного корпуса текстов *Cf* и тематического *Ct* можно считать, что

$$P(S_{Ct}/H_{Ct}) \approx \frac{n_{Ct}}{N_{Ct}}, \quad P(S_{Cf}) \approx \frac{n_{Cf}}{N_{Cf}}, \quad P(H_{Ct}) \approx \frac{N_{Ct}}{N_{Cf}},$$

где *nCt*, *nCf* – абсолютные частоты встречаемости словоформы α в тематическом и полном корпусах текстов, а *NCt*, *NCf* – число вхождений α в корпусы текстов *Ct* и *Cf* соответственно. Тогда формула для вычисления информативности I^{α}_{Ct} словоформы α в тематическом корпусе текстов Ct примет вид

$$I^{\alpha}_{Ct} = \frac{n_{Ct}}{n_{Cf}}. \qquad (1)$$

Для вычисления информативности словоформ в текстовых документах введем понятие отношения информативности.

Пусть α – некоторая словоформа, а P_{Cf} и P_{Ct_i} ($i = \overline{1, n}$) – ее абсолютные частоты соответственно в полном и i-м тематическом корпусах текстов. Тогда *(n + 2)*-арное отношение

$$\Omega = \{(\alpha, P_{Cf}, P_{Ct_1}, P_{Ct_2}, \ldots, P_{Ct_n}) \mid \alpha \in Cf\}$$

будем называть отношением информативности.

2.3. Словари базы знаний, используемые при индексировании

Сформируем на основе отношения информативности Ω совокупность *W* кортежей типа $\langle \alpha, P_{Cf}, P_{Ct_1}, P_{Ct_2}, \ldots, P_{Ct_n} \rangle$, где α – произвольная словоформа полного корпуса текстов *Cf*, которую будем назвать частотным словарем словоформ. Данный словарь предназначен для нахождения статистических характеристик словоформ из индексируемого документа в полном и тематических корпусах текстов.

Пусть αj – произвольный поисковый признак из словаря *W*. Тогда совокупность *WPr* кортежей типа $\langle \alpha 0, \alpha j \rangle$, где $\alpha 0$ – ключевые слово (или код) парадигмы словоформы αj, будем называть словарем словоизменительных парадигм. Данный словарь служит для поиска всех словоизменений αj после ее нахождения в словаре *W*.

Словарь синонимичных словоформ *WSyn* представляет собой совокупность кортежей типа $\langle \alpha, \beta, \gamma, \ldots \rangle$, где α, β, γ – произвольные словоформы из словаря WPr, связанные отношением синонимии. Данный словарь позволяет по коду парадигмы для определенной словоформы найти его синонимы.

Словари *WPr* и *WSyn* формируются в человеко-машинном режиме с использованием специализированных программных средств, однако на первоначальном этапе подсистема поиска может функционировать без данных словарей, что приведет к некоторому ухудшению качества индексирования и последующего поиска текстовых документов, но не скажется на ее работе в целом.

Словарь *W* является единым для всех тематических разделов предметной области нанотехнологий, т. е. в него попадают словоформы из всех текстовых документов, которые добавляются в полный корпус *Cf*. Данный словарь создается программно.

2.4. Алгоритм индексирования текстовых документов

Представим поисковый образ *Ot* некоторого текстового документа *t* в виде:

$$Ot = \{(a, Ia) \mid a \in W,\ 0 \le Ia \le 1\}$$

где *Ia* – информативность словоформы *a*.

Алгоритм индексирования, формирующий поисковые образы *Ot* полнотекстовых документов, работает следующим способом: выбирается очередная словоформа a документа *t* и по формуле (1) вычисляется ее информативность с учетом словоизменений и синонимии. Для исключения из поискового образа повторений информации позиции обработанных словоформ и их словоизменений в документе *t* запоминаются. Формируется первая пара *(a, Ia)* множества *Ot*. После обработки всех словоформ документа t создается его поисковый образ *Ot*.

Алгоритм 1. На входе алгоритма – полнотекстовый документ t, словари базы знаний *W, WPr, WSyn* и полный корпус текстов *Cf*, на выходе – поисковый образ *Ot* документа *t*. Алго-

ритм индексирования включает в себя следующие шаги:

1. *Ot := ∅.*
2. Выбрать очередную словоформу a из текста t.
3. Если *a* или ее словоизменения были до этого обработаны, то перейти к п. 2, иначе – к п. 4.
4. Найти a в словаре *W* и получить для нее код парадигмы.
5. В словаре *WPr* найти все словоизменения словоформы a.
6. Найти все синонимичные словоформы для a в словаре *WSyn*.
7. Определить количество вхождений a в документе t – n_1^t.
8. Определить количество словоизменений словоформы в документе – n_2^t.
9. Запомнить позиции a и ее словоизменений в документе t.
10. Определить количество вхождений в документе t словоформ, которые являются синонимами a, – n_3^t.
11. Определить на основе статистической информации, полученной из словаря W, суммарное число вхождений словоформы, его словоизменений и синонимов в Cf – N_{Cf}^1, N_{Cf}^2 и N_{Cf}^3.
12. По формуле (1) вычислить значение информативности *Ia* текущей словоформы *a*, определяя абсолютные частоты встречаемости, как сумму соответствующих значений (*nCt* = $= n_1^t + n_2^t + n_3^t$ и nCf = $N_{Cf}^1 + N_{Cf}^2 + N_{Cf}^3$).
13. Поместить пару (a, Ia) в множество Ot.
14. Если все словоформы документа *t* исчерпаны, то *КОНЕЦ* (поисковый образ *Ot* текста *t* сформирован), иначе перейти к п. 2.

Трудоемкость алгоритма *1* – не более чем *l + 1* повторений всех его шагов, где *l* – количество словоформ в полнотекстовом документе *t*.

3. Поиск текстовых документов

Представим первоначальный запрос пользователя на поиск информации в виде множества *z1 =((b1, 1), (b2, 1), ...)*, где *bi* ($i = \overline{1, n}$) – словоформы *z1*. Для достижения адекватности представления информационной потребности пользователя в первоначальном запросе z1 может потребоваться его дальнейшая коррекция на основе соответствующего ему динамического корпуса текстов Dz1.

Под уточненным поисковым предписанием z_{Dz_1}, учитывающим информацию из динамического корпуса текстов *Dz1*, будем понимать множество $z_{Dz_1} = \{(\text{bj}, J_{b_j}) \mid j = \overline{1, k}\}$, где *b1, b2,..., bk* – словоформы корпуса *Dz1*, J_{b_j} – информативности соответствующих словоформ, вычисленные по формуле (1), такое, что все словоформы z1 принадлежат z_{Dz_1}.

Для достижения высоких значений полноты и точности целесообразно включать в множество z_{Dz_1}, помимо словоформ *z1*, словоформы динамического корпуса текстов *Dz1*, информативность которых больше некоторого порогового значения *J0* (данный параметр может быть получен экспериментальным путем или задаваться пользователем в процессе поиска).

Процедура коррекции первоначальных запросов пользователей может быть осуществлена с использованием следующих стратегий [6]:

– Документы из динамического корпуса текстов Dz1 предъявляются пользователю, который исключает из него все непертинентные тексты. Полученное в результате множество (обозначим его через *Dz2*) считается уточненным динамическим корпусом, на основе которого путем его индексирования формируется уточненное поисковое предписание z_{Dz_1}. Каждой словоформе ставится в соответствие ее информативность в корпусе *Dz2*.

– Множество документов из динамического корпуса текстов *Dz1* без предъявления пользователю считается уточненным динамическим корпусом текстов, на основе которого путем его индексирования формируется уточненное поисковое предписание z_{Dz_1} в виде совокупности словоформ и соответствующих значений информативности в *Dz1*.

3.1. Критерий выдачи при ранжировании найденных документов

Представим поисковый образ документа Ot в векторном виде следующим образом.

Введем в рассмотрение n-мерное евклидово пространство. Для этого лексикографически упорядочим все словоформы соответствующего документа, т. е. сформируем кортеж *Pr = ⟨a1, a2,..., an⟩.* Для индексируемого текстового документа t построим вектор в евклидовом пространстве:

Ot = (I_{a_1}, I_{a_2},…, I_{a_n}).

Координатами вектора *Ot* являются значения информативности соответствующих словоформ.

Аналогичным образом построим векторное представление уточненного поискового предписания z_{Dz_1} в виде $\mathbf{F}_{z_{Dz_1}} = (J_{b_1}, J_{b_2}, \ldots, J_{b_k})$, где J_{b_j} – информативность словоформы bj в z_{Dz_1}.

Тогда мера близости z_{Dz_1} и *Ot* представляется в виде

$$\cos\varphi = \frac{\mathbf{O}_t \, \mathbf{F}_{z_{Dz_1}}}{|\mathbf{O}_t|\,|\mathbf{F}_{z_{Dz_1}}|} = \frac{\sum_{i,j=1}^{n,k} I_{a_i} J_{b_j}}{\sqrt{\sum_{i=1}^{n} I_{a_i}^{\,2}}\sqrt{\sum_{j=1}^{k} J_{b_j}^{\,2}}}. \qquad (2)$$

3.2. Алгоритм поиска текстовых документов

Рассмотрим алгоритм информационного поиска текстовых документов, который реализует описанную выше стратегию и использует критерий выдачи (2).

Алгоритм 2. На входе алгоритма – множество Твх текстовых документов, множество $O_{T_{\text{âõ.}}}$ их поисковых образов, запрос пользователя *z1* $\in$ *Zвх*, словари базы знаний *W, WPr, WSyn*, полный корпус текстов *Cf* и пороговое значение информативности *J0*, на выходе – кортеж найденных и ранжированных текстовых документов. Алгоритм включает в себя следующие шаги:

1. *M* := ∅, *K* := ∅.

2. Сформировать динамический корпус текстов *Dz1* и получить уточненное поисковое предписание пользователя z_{Dz_1} по запросу z1.

3. Найти в множестве $O_{T_{\text{âõ}}}$ поисковые образы текстовых документов, для которых cos φ ≠ 0 (формула (2)). Поместить результаты поиска в множество M.

4. Если M := ∅, то перейти к п. 6, иначе – к п. 5.

5. Упорядочить все поисковые образы документов множества *M* по убыванию значений критерия выдачи (2) и поместить соответствующие им тексты из множества *Твх* в кортеж $K = \langle t_1, t_2, \ldots, t_r \rangle$.

6. *КОНЕЦ* (кортеж *K* найденных и ранжированных документов сформирован).

Алгоритм 2 заканчивает работу не более чем за один проход всех его шагов.

4. Реферирование текстовых документов

4.1. Словари базы знаний, используемые при реферировании

При реферировании текстовых документов используются тематические корпусы текстов, а также словарь прагматически полных синтагматических структур (ПП-структур) и ситуативный словарь.

ПП-структура – это информативная в некотором тематическом разделе предметной области (т. е. хотя бы в одном тематическом корпусе текстов) синтагматическая структура, выражаемая устойчивым словосочетанием. Примеры ПП-структур: информационная технология; гидрофизический институт; углеродные нанотрубки; молекулярные нанотехнологии; синтаксический анализ предложения.

Ситуативный словарь. Словарь формируется на основе ситуативного отношения, которое определим следующим образом.

Пусть *Ctj* ($i = \overline{1, n}$; n ≥ 2) – тематические корпусы текстов, *Cf* *(Cf = Ct1 ∪ Ct2 ∪ Ctn)* – полный корпус текстов, а *Wo* множество всех слов полного корпуса текстов *Cf*. Тогда отношение толерантности *Θ* (рефлексивное и симметричное бинарное отношение) на множестве *Wo* назовем ситуативным отношением в полном корпусе текстов *Cf*, если любая упорядоченная пара информативных *слов (a, b)* из множества *Wo* является элементом отношения *Θ* тогда и только тогда, когда вероятность совместной встречаемости слов *a* и *b* в корпусе текстов *Cf* не меньше некоторого порогового значения. (Эту вероятность будем называть информативностью ситуативной связи слов.) Под совместной встречаемостью двух слов здесь понимается их наличие (а также их синонимов и словоизменений) в одном и том же предложении корпуса *Cf*. Граф *Scum*. ситуативного отношения, каждое ребро которого помечено значением информативности соответствующей ситуативной связи, будем называть ситуативной сетью. Сеть *Scum*. реализуется в виде ситуативного словаря. В первых двух столбцах словаря содержатся пары слов, а в третьем – информативность ситуативной связи этих слов в процентах.

Рассмотренные словари создаются в полуавтоматическом режиме с использованием программно сформированных исходных файлов и средств визуализации АРМ эксперта-лингвиста.

4.2 Построение реферата

При построении реферата текстового документа будем использовать граф информативности текста и граф ситуативных связей между его предложениями [7, 8].

Пусть имеется текст (т. е. кортеж предложений) $T = \langle \pi 1, \pi 2, \ldots \rangle$. Вычислим информативность всех предложений текста T и исключим из T неинформативные предложения, т. е. все предложения π, информативность $I\pi$ которых меньше некоторого $I0$. В результате получим кортеж предложений $Тинф. = \langle \pi_{i_1}, \pi_{i_2}, \ldots \rangle$, который будем называть маршрутом информативности текста T. Соединив последовательно вершины графа текста GT (т. е. графа редукции линейного порядка на множестве всех предложений текста T), соответствующие информативным предложениям, получим орграф Gинф., который будем называть графом информативности текста T.

Пусть $T+$ – множество всех информативных, а $T-$ – всех неинформативных предложений текста T ($T = T+ \cup T-$). Определим на паре множеств T +, T – симметричное отношение Ξ, такое, что для любых предложений $\pi \in T+$ и $\rho \in T-$ $(\pi, \rho) \in \Xi$ тогда и только тогда, когда информативность $I\pi\rho$ ситуативной связи между предложениями π и ρ не меньше некоторого значения. Граф отношения Ξ назовем графом ситуативных связей между предложениями текста T.

4.3 Алгоритм выявления информативных предложений

Алгоритм 3. На входе алгоритма – текст T объемом в N предложений, на выходе – текст T с выделенными информативными предложениями.

1. Вычислить информативность Ia каждой словоформы a текста T по формуле

$$I_a = \frac{n_T}{n_{Cf}},$$

где nT , nCf – абсолютные частоты встречаемости словоформы a в тексте T и полном корпусе текстов Cf соответственно.

2. Вычислить информативность $I\pi$ каждого предложения π текста T по формуле

$$I_\pi = \frac{I_a + I_b + \ldots}{\sqrt{I_a{}^2 + I_b{}^2 + \ldots}},$$

где $Ia, Ib, \ldots$ – значения информативности всех слов предложения π.

3. Упорядочить все N предложений текста T по убыванию их информативности.

4. Выделить из полученного упорядоченного списка n наиболее информативных предложений (n задается в качестве параметра).

5. Восстановить исходный порядок предложений в тексте T. Конец.

В соответствии с алгоритмом 3 в тексте выявляются информативные предложения, т. е. строится маршрут информативности $Тинф.$.

4.4 Алгоритм построения контекста информативных предложений

Алгоритм 4. На входе алгоритма – текст T с выделенными информативными предложениями, на выходе – контекст информативных предложений $E1, E2, \ldots, En$.

1. Ei := ∅, $i = \overline{1, n}$ (i – номер информативного предложения текста T).

2. i := 1.

3. π := πi.

4. Вычислить значения информативности ситуативных связей Iπσ для всех неинформативных предложений σ, предшествующих предложению π и следующих за ним, по формуле

$$I_{\pi\sigma} = \frac{I_{ab} + I_{cd} + \ldots}{\sqrt{I_{ab}{}^2 + I_{cd}{}^2 + \ldots}}.$$

5. Поместить все предложения σ, для которых $I\pi\sigma$ превышает пороговый уровень $I^0_{\pi\sigma}$ (задается в качестве параметра) в множество Ei (в порядке их появления в тексте T).

6. $i := i + 1$. Если $i \le n$, то перейти к п. 4, иначе – конец (контекст информативных предложений построен).

5. Программная реализация электронного банка данных по нанотехнологиям

Подсистемы электронного банка данных по нанотехнологиям представлены в виде комплекса программного обеспечения, разработанного на языке программирования *C++ c* использованием кроссплатформенной библиотеки *Qt*. Тематические корпусы текстов и лингвистические словари базы знаний хранятся в *СУБД MySQL*, которая предоставляет возможность организации многопользовательского режима работы, а также реализует технологию клиент-сервер. В этом режиме запросы пользователей выполняются на стороне сервера, что позволяет обеспечивать высокую производительность и эффективное совместное применение ресурсов баз данных.

Заключение

В статье представлены результаты проекта по созданию электронного банка данных по

нанотехнологиям. Разработанный банк данных позволит оперативно получать сведения о состоянии и возможностях применения нанотехнологических исследований в науке и технике, вести автоматическую классификацию, накопление и анализ текстовых документов большого объема. Программное обеспечение банка данных может быть использовано в научно-технических библиотеках, в информационно-аналитических отделах различных служб и организаций, которые осуществляют оперативный сбор и аналитическую обработку текстовых документов по различным предметным областям в Интернете, локальных сетях и на жестких или съемных дисках отдельных компьютеров.

Список литературы

1. Ландэ, Д.В. Интернетика. Навигация в сложных сетях: модели и алгоритмы / Д.В. Ландэ, А.А. Снарский, И.В. Безсуднов. – М. : Либроком (Editorial URSS), 2009. – 264 с.
2. Manning, C. Introduction to Information Retrieval / C. Manning, P. Ragha-van, H. Schütze. – 1 edition. – Cambridge University Press, 2008. – 496 p.
3. Тактаев, С. Поиск информации в компьютерных сетях: новые подходы / С. Тактаев // [Электронный ресурс]. – Режим доступа: http://www.searchengines.ru/articles/004603.html. – Дата доступа: 15.02.2012.
4. Мамчич, А.А. Алгоритмы индексирования и поиска документов на основе динамических корпусов текстов / А.А. Мамчич // Информатика. – 2010. – № 1. – С. 82–90.
5. Липницкий, С.Ф. Веб-поиск и аннотирование научно-технической информации на основе тематических корпусов текстов / С.Ф Липницкий, А.А. Мамчич, С.А. Сорудейкина // Информатика. – 2009. – № 2. – С. 114–126.
6. Липницкий, С.Ф. Моделирование информационного поиска на основе динамических корпусов текстов / С.Ф. Липницкий, А.А. Мамчич // Весці НАН Беларусі. Сер. фіз.-тэхн. навук. – 2011. – № 1. – С. 72–81.
7. Кравцов, А.А. Система автоматического индексирования и реферирования текстовых документов / А.А. Кравцов, С.Ф. Липницкий, Л.В. Степура // Таврический вестник информатики и математики. – 2008. – № 1. – С. 260–266.
8. Степура, Л.В. Алгоритмы построения контекста информативных предложений при реферировании текстовых документов / Л.В. Степура // Информатика. – 2010. - №2. - С. 46-53.

Сведения об авторах

Липницкий Станислав Феликсович – доктор технических наук, доцент, главный научный сотрудник Объединенного института проблем информатики НАН Беларуси.
E-mail: lipn@newman.bas-net.by

Мамчич Алексей Анатольевич – кандидат технических наук, научный сотрудник Объединенного института проблем информатики НАН Беларуси.
E-mail: lexamam@newman.bas-net.by

Степура Людмила Васильевна – научный сотрудник Объединенного института проблем информатики НАН Беларуси
E-mail: stepura@newman.bas-net.by

Статья поступила в редакцию 20.08.2011

