УДК 004.9

Д. С. Замятин *, А. Ю. Михайлюк **, кандидаты техн. наук,
Е. С. Михайлюк *, А. В. Петрашенко *, канд. техн. наук,
А. В. Пилипчук *, аспирант, **В. П. Тарасенко ***, д-р техн. наук
* Национальный технический университет Украины
«Киевский политехнический ин-т»
(Украина, 03056, Киев, пр-т Победы, 37,
тел. (044)4549031, (044) 4549492, (044) 4068476, (044) 2363202;
E-mail: dsz@ukr.net, petrashenko@gmail.com, mes@scs.ntu-kpi.kiev.ua,
ilexcorp@ukr.net, vtarasen@scs.ntu-kpi.kiev.ua)
** Киевский университет им. Б. Гринченко
(Украина, 04212, Киев, ул. Тимошенко, 13-Б,
тел. (044) 4268405, E-mail: may-62@ukr.net)

Квазисемантический поиск текстовых данных. Способы модификации запроса

(Статью представил д-р техн. наук В. Я. Кондращенко)

Предложены формальные модели процедур внесения семантики в поисковый запрос при реализации квазисемантического поиска. Модификация запроса выполнена с использованием лингвистической онтологии предметной области. Рассмотрена специфика отображения поискового запроса из поля терминов в поле синсетов онтологии. Предложены и исследованы три стратегии коррекции поискового запроса: горизонтальная, вертикальная и ассоциативная. Проанализировано влияние процедур модификации поискового запроса на поисковый отклик.

Запропоновано формальні моделі процедур внесення семантики у пошуковий запит при реалізації квазісемантичного пошуку. Модифікацію запиту виконано з використанням лінгвістичної онтології предметної галузі. Розглянуто специфіку відображення пошукового запиту із поля термінів в поле синсетів онтології. Запропоновано та досліджено три стратегії корекції пошукового запиту: горизонтальна, вертикальна та асоціативна. Проаналізовано вплив процедур модифікації пошукового запиту на пошуковий відгук.

Ключевые слова: поиск, квазисемантика, лингвистическая онтология, модификация запроса.

В условиях постиндустриального общества необходимым условием практически любой общественно-полезной деятельности становится использование высокоэффективных средств поиска необходимой информации в глобальном электронном информационном пространстве. В настоящее время

для обеспечения информационных потребностей пользователей всех категорий используются в основном традиционные полнотекстовые поисковые машины [1], широко представленные в глобальной сети Internet.

Отдавая должное весомому вкладу таких поисковых систем в процесс усовершенствования информационной деятельности, необходимо отметить их основной недостаток, а именно отсутствие возможности непосредственного учета семантики в поисковом процессе.

В тоже время, следует заметить, что относительно перспектив реального прорыва в области информационного поиска классическая концепция полнотекстового поиска практически себя исчерпала [2, 3]. В настоящее время необходимо использование высокоинтеллектуальных методов доступа к распределенным по глобальной компьютерной сети данным и знаниям [4, 5]. Однако развитие собственно семантических машин происходит недостаточно быстро в связи со значительной сложностью практической реализации эффективного семантического анализа естественно-языковых текстов [6].

Именно поэтому существенный теоретический и практический интерес представляет концепция квазисемантической поисковой системы, согласно которой привнесение смысловой компоненты в процесс поиска осуществляется пользователем собственноручно посредством итерационной модификации начальной версии запроса с использованием интеллектуального редактора запроса, построенного на основе лингвистической онтологии предметной области [7]. Реализация процесса функционирования такого редактора, по аналогии с экспертной системой [8], позволяет анализировать содержание, фактически заложенное пользователем в текущий поисковый запрос, и предлагать на выбор такую модификацию последнего, которая бы в конечном итоге позволяла найти информационные объекты, максимально отвечающие настоящему поисковому интересу пользователя.

Согласно упомянутой концепции, для формирования и модификации поискового запроса пользователя могут быть применены три следующие основные процедуры, в которых используется логическое отображение поискового запроса в поле понятий онтологии.

1. *Процедура горизонтальной коррекции запроса.* Выполняется в три этапа: поиск понятий онтологии, соответствующих терминам запроса; разрешение неоднозначности многозначных терминов при участии пользователя на основе определений соответствующих понятий; расширение запроса за счет синонимических рядов каждого из понятий.

Эта процедура позволяет представить поисковый запрос, изначально заданный терминами естественного языка, в виде понятий, имеющих

вполне определенное значение, адекватное поисковым интересам пользователя. В последующих процедурах используется именно такое понятийное представление поискового запроса. При необходимости такое представление можно легко транслировать в форму терминов, уже более точно учитывая смысл определяемых ими понятий.

2. *Процедура вертикальной коррекции поискового запроса.* Характерной чертой любой онтологии является ее иерархическая упорядоченность [9]. Понятия объединяются в общую структуру, представляющую собой дерево, в корне которого находится наиболее общее понятие. Чем ниже спускаться по иерархии онтологии, тем более конкретными будут понятия. Используя эту особенность, можно предложить модификацию поискового запроса посредством навигации по вертикальным связям онтологии и включения в запрос понятий, которые логически связаны с терминами первоначальной версии запроса, но более или менее конкретными в зависимости от информационного интереса пользователя. Таким образом можно получить соответственно более конкретные или более общие результаты в поисковой выборке по сравнению с поисковым отзывом на немодифицированный поисковый запрос.

3. *Процедура ассоциативной коррекции поискового запроса* из всех понятий онтологии позволяет выбрать для модификации поискового запроса лишь те, которые потенциально близки по смыслу к поисковому запросу с учетом контекста. Такие понятия будут составлять своеобразный семантический срез онтологии по поисковому запросу. Связи онтологии в данном случае рассматриваются как ассоциации между объектами, а сама онтология — как многомерная семантическая сеть. Найденный семантический срез может быть выборочно использован пользователем для модификации поискового запроса с учетом возможного уточнения информационного интереса.

Предлагаемая формализация ключевых процедур квазисемантического поиска текстовых данных направлена на обеспечение высокой эффективности программных реализаций. Исследуем такие задачи:

- формализация описания информационной инфраструктуры, на которой основан квазисемантический поиск текстовых данных;
- разработка математической интерпретации процедур формирования, коррекции и реализации поискового запроса в процессе квазисемантического поиска;
- исследование влияния процедур коррекции поискового запроса на результы поисковой выборки.

В ходе исследования будем считать, что в общем случае поисковый индекс текстовых информационных объектов может быть как классичес-

ким, т.е. организованным на основе терминов естественного языка [10], так и семантически-ориентированным или концепт-индексом, т.е. построенным на основе понятий [11]. Поэтому на формально-логическом уровне будем рассматривать и сравнивать реализации квазисемантического поиска в текстоориентированных базах и хранилищах данных, проиндексированных согласно обоим упомянутым подходам. При этом особенности программно-технического воплощения поиска информационных объектов подробно не рассматриваются, так как эти исследования должны быть проведены в рамках отдельной задачи построения соответствующей поисковой машины.

Особенности инфраструктуры квазисемантического поиска. Для анализа всех процедур квазисемантического поиска необходимо ввести формальное описание данных, с которыми эти процедуры взаимодействуют. Поэтому следует определить основные понятия: служебный лингвистический ресурс, используемый в алгоритмах поиска, множество текстовых информационных объектов (документов) и поисковый запрос.

В качестве служебного лингвистического ресурса для реализации квазисемантического поиска будем использовать лингвистическую онтологию. Согласно [9] лингвистический подход к формированию и исследованию баз знаний онтологического типа основан на изучении значительных массивов (корпусов) естественного языка. Получаемая таким образом лингвистическая онтология проецирует соответствующую область знаний, с одной стороны, в естественно-языковое поле, а с другой — в систему семантических отношений.

Согласно [12] онтология может быть представлена тройкой $O = =< X, R, \Phi >$, где X — конечное множество концептов (понятий). В контексте рассмотрения процедур квазисемантического поиска под множеством концептов X будем понимать конечное множество синсетов $S = \{s_i : i = \overline{1, I}\}$, где I — число синсетов в онтологии.

Синсет — это уникальная смысловая единица, интегрирующая в себе набор близких синонимов, выраженных терминами (словами или словосочетаниями) естественного языка, а также однозначное определение, позволяющее отличить его от других синсетов. Таким образом, каждый синсет s описывается двойкой

$$s =< T^s,\ D >, \tag{1}$$

где D — дескриптор — описательное представление понятия, выражаемого синсетом; T^s — конечное непустое множество терминов-синонимов, составляющих синсет. Каждый синсет $s \in S$ отображается в поле терминов

$$T^s = \{t_0, t : t_0 \in T,\ t \in T,\ Syn(t_0, t)\}, \tag{2}$$

где T — множество терминов естественного языка; t_0 — основной термин синсета (обычно определяется как наиболее употребляемый из определяющих соответствующее понятие); t — остальные термины, определяющие понятие s; Syn — функция синонимии двух терминов. Далее под синсетом s, если не указано иного, будем понимать его отображение в поле терминов T^s. В общем случае $s_i \cap s_j \neq \varnothing$ при $i \neq j$, поскольку вследствие омонимии произвольный термин может одновременно принадлежать синсетам s_i и s_j.

Пусть R — конечное множество семантических отношений, в состав которого могут входить семантические отношения любых видов. Однако в общем случае для задачи квазисемантического поиска будем различать два их вида: $R = \{r_h, r_a\}$, где r_h — отношение гиперонимии, а r_a — отношение ассоциации. Отношение r_h в связке с множеством синсетов в данном случае определяет таксономию понятий онтологии.

Пусть Ф — конечное множество функций интерпретации, содержание которого в контексте проблематики квазисемантического поиска является функцией интерпретации отношений онтологии, определяющей вес семантического отношения w между синсетами. Чем больше значение веса, тем сильнее выражена соответствующая семантическая связь между двумя понятиями.

Выбор данного типа онтологии в качестве служебного лингвистического ресурса определен следующими соображениями:

- эффективность лингвистической онтологии в качестве универсального средства описания понятийного аппарата различных предметных областей;
- широкая распространенность подобных онтологий (например, англоязычная лингвистическая онтология WordNet и ее локализации для других естественных языков [9]) и возможность приведения к данному виду других лингвистических баз знаний онтологического типа;
- сравнительная простота автоматической машинной обработки лингвистической онтологии в случае использования последней на различных этапах информационного поиска;
- потенциальная возможность автоматизации процедур синтеза лингвистической онтологии (или наращивания уже существующей) вследствие применения программного инструментария интеллектуального анализа текстовых массивов.

Множество документов рассматривается как конечное подмножество коллекции естественно-языковых текстовых информационных объектов $\{d_j : j = \overline{1, J}\} \subseteq D$, где J — число документов в коллекции. Каждый доку-

мент d из коллекции текстовых информационных объектов D классически индексируется конечным множеством терминов $\{t_g : g = \overline{1,G}\} \subset T$, где G — число терминов, с помощью которых индексируется документ. Однако квазисемантический подход к поиску текстовых информационных объектов предусматривает также возможность индексации множеством синсетов $\{s_i : i = \overline{1,I_d}\} \subset S$, где I_d — число синсетов, с помощью которых индексируется документ.

В первом случае конечное множество терминов естественного языка, описывающих (индексирующих) документ d,

$$T^d = \{t : t \in T, F^t(t) > 0\},\ T^d \subset T^D, \tag{3}$$

где T^D — множество терминов полнотекстового индекса общей коллекции документов D; $F^t(t)$ — функция присутствия термина в индексируемом документе. В самом простом случае эта функция принимает ненулевое значение, когда термин t присутствует в документе: $\{t\} \cap T^d \neq \varnothing$, где $t \in T$, т.е. $\exists\, t \in T : t \in T^d$.

Во втором случае конечное множество синсетов онтологии, описывающих (индексирующих) документ d,

$$S^d = \{s : s \in S, F^s(s) > 0\},\ S^d \subset S^D, \tag{4}$$

где S^D — множество синсетов концепт-индекса полной коллекции документов D; $F^s(s)$ — функция семантической принадлежности синсета индексируемому документу. В самом простом случае эта функция принимает ненулевое значение, если $s \cap T^d \neq \varnothing$, где $s = \{t : t \in T\}$, т.е. $\exists\, t \in T : t \in T^d, t \in s$.

Поисковый запрос — это формализованный способ выражения информационного интереса пользователя. Поэтому чаще всего в начальной форме поисковый запрос выполняется на естественном языке. В общем случае поисковый запрос пользователя, выраженный на конечном множестве терминов естественного языка, можно записать в виде $Q_t = \{t : t \in T\}$.

Большинство текстовых поисковых систем позволяют подавать поисковый запрос в виде логического выражения. Используемые при этом основные логические операторы — *AND*, *OR*, *NOT*.

Запрос вида $Q = t_i\ AND\ t_j$, $i \neq j$, в результате поиска выдает лишь те документы $\{d_k : k = \overline{1,K}\} \subseteq D$, где K — число выбранных документов, которые содержат термин t_i и термин t_j одновременно.

Запрос вида $Q = t_i\ OR\ t_j$, $i \neq j$, в результате поиска выдает объединение выборок документов по термину t_i и термину t_j.

Запрос вида $Q = NOT\ t_i$ в результате поиска выдает лишь те документы $\{d_k : k = \overline{1, K}\} \subseteq D$, где K — число выбранных документов, которые не содержат t_i. Однако фактически итоговая выборка документов по такому поисковому запросу может быть получена вследствие реализации дополнения множества документов, содержащих термин t_i, к общей коллекции документов D.

Учитывая тот факт, что в большинстве случаев поисковый запрос формулируется на естественном языке и в соответствии с классическим синтаксисом запросов поисковых систем рассматривается как термины, объединенные оператором *AND* [13], поисковый запрос далее будем считать множеством терминов естественного языка, каждый из которых должен входить в искомый текстовый информационный объект.

В случае более сложного запроса с использованием операторов *OR* или *NOT* весь запрос может быть разбит на подзапросы с последующим объединением результатов или исключением лишних в соответствии с семантикой конкретного логического оператора. Поэтому для упрощения формального представления процедур квазисемантического поиска далее запрос будем рассматривать как набор терминов, связанных оператором *AND*. В случае необходимости все приведенные выкладки формально могут быть получены для любого варианта сложного поискового запроса.

Формализация процедур проекции поискового запроса на онтологию и горизонтальной коррекции запроса. Процедуры квазисемантического поиска работают не с терминами, а с понятиями, или в контексте онтологии — с синсетами. Поэтому для их работы вначале нужно связать исходный поисковый запрос Q^t с онтологией. Введем функцию отображения терминов запроса в множество синсетов

$$\varphi(t) : Q^i \to Q^s, \tag{5}$$

где $Q^s = \{s : s \in S\}$ — поисковый запрос, выраженный конечным множеством синсетов онтологии предметной области; $\varphi(t)$ — отображение множества терминов запроса в множество синсетов. Правило отображения можно сформулировать так:

$$\forall\ t \in Q^t \to s \in S : t \in s.$$

Запрос в виде Q^s позволит выполнять поиск результатов в случаях:

а) когда документы коллекции D описываются концепт-индексом S^D, — определив функцию подобия документа и запроса как пересечение множеств $Q^s \cap S^d$, т.е. на уровне концептов, а не на уровне терминов;

б) когда документы коллекции D описываются множеством T^D, — используя не начальный запрос Q^t, а модифицированный Q^{t+}, полученный с помощью преобразования Q^s:

$$Q^{t+} = \bigcup_j (s_j : s_j \in Q^s) = \bigcup_j \left(\bigcup_i (t_i : t_i \in s_j) : s_j \in Q^s \right). \quad (6)$$

Таким образом, запрос Q^{t^+} в соответствии с (2) автоматически расширяется посредством синонимических связей (возможно, с последующей селекцией пользователем), а поиск документов происходит с использованием функции подобия, в основе которой лежит пересечение множеств терминов модифицированного запроса и множества терминов каждого документа $Q^{t+} \cap T^d$.

Для определения влияния процедуры связывания поискового запроса с онтологией на результаты поиска по классическому полнотекстовому индексу (когда процедура поиска функционирует на уровне терминов) рассмотрим следующую теорему.

Теорема 1. Поисковый запрос Q^{t+}, полученный с использованием синонимических связей онтологии предметной области, во время поиска в термин-ориентированном индексе увеличивает полноту поискового отклика по сравнению с немодифицированным запросом Q^t.

Доказательство. Пусть начальный запрос состоит из одного термина $Q^t = \{t_q\}$, где $t_q \in T$. Тогда поисковый отклик формируется на основе пересечения индекса коллекции документов T^d с запросом:

$$T^d \cap Q^t = T^d \cap \{t_q\}.$$

Обозначим D_1 множество документов поискового отклика по термину t_q, а $D_1^{\text{рел}} \subseteq D_1$ — множество релевантных документов в отклике. Согласно [14] полноту поискового отклика рассчитываем по формуле

$$\Pi = \frac{|D_{\text{п.рел}}|}{|D_{\text{общ.рел}}|},$$

где $D_{\text{п.рел}}$ — множество полученных в поисковом отклике релевантных документов; $D_{\text{общ.рел}}$ — общее множество релевантных документов в коллекции.

Для начального запроса Q^t полнота поискового отклика имеет вид

$$\Pi_1 = \frac{|D_1^{\text{рел}}|}{|D_{\text{общ.рел}}|}.$$

После преобразования (5) получаем $Q^s = \{s_q\}$. Здесь $s_q \in S$ и $s_q = \{t_0, t_1, ..., t_q, ..., t_n\}$ в соответствии с (2), где n — число синонимов t_0 в синонимическом ряду синсета s_q. Исходя из (6) находим

$$Q^{t+} = \bigcup_j (s_j : s_j \in Q^s) = \bigcup_j \left(\bigcup_i (t_i : t_i \in s_j) : s_j \in Q^s \right) =$$
$$= \bigcup_i (t_i : t_i \in s_q) = (t_0 \cup t_1 \cup ... \cup t_q \cup ... \cup t_n).$$

Поисковый отклик формируется на основе пересечения индекса коллекции с запросом:

$$Q^{t+} \cap T^D = (t_0 \cup t_1 \cup ... \cup t_q \cup ... \cup t_n) \cap T^D =$$
$$= (t_0 \cap T^D) \cup (t_1 \cap T^D) \cup ... \cup (t_q \cap T^D) \cup ... \cup (t_n \cap T^D).$$

Каждое выражение $(t_i \cap T^D)$ определяет некоторую выборку документов, назовем ее D_i, а соответствующие множества — выборку релевантных документов $D_i^{\text{рел}} \subseteq D_i$. Для модифицированного запроса Q^{t+} полнота имеет вид

$$\Pi_2 = \frac{\left| D_0^{\text{рел}} \cup D_1^{\text{рел}} \cup ... \cup D_q^{\text{рел}} \cup ... \cup D_n^{\text{рел}} \right|}{\left| D_{\text{общ.рел}} \right|}.$$

Поскольку все термины $t_0, t_1, ..., t_q, ..., t_n$ являются синонимами, общее количество релевантных документов по поисковому запросу равно мощности объединенного множества релевантных документов по каждому из терминов.

Очевидно, что $\Pi_2 \geq \Pi_1$. Следовательно, полнота поискового отклика по модифицированному поисковому запросу в общем случае возрастает по сравнению с полнотой поискового запроса по немодифицированному запросу, что и требовалось доказать. Результат доказательства теоремы несложно расширить на запрос, состоящий из m слов, где $m \geq 1$.

В процедуре (5) с учетом (2) возможна ситуация, когда $t \in Q^t$ удовлетворяет двум и более неравенствам одновременно

$$t \cap s_i \neq \varnothing,$$
$$t \cap s_j \neq \varnothing,$$
$$...$$
$$t \cap s_k \neq \varnothing, \qquad (7)$$

где $i \neq j \neq ... \neq k$.

Теорема 2. Если в начальном поисковом запросе Q^t присутствует термин, имеющий несколько значений, то без разрешения проблемы омонимии при отображении множества терминов запроса во множество синсетов точность поисковой выборки документов по запросу Q^{t+} уменьшается.

Доказательство. Пусть $Q^t = \{t^b\}$, где $t^b \in T$ — многозначный термин, т.е. верным является выражение (7). Тогда в соответствии с правилом отображения (5)

$$Q^s = \{s_i, s_j, ..., s_k\}, \; i \neq j \neq ... \neq k .$$

Таким образом, модифицированный поисковый запрос для поиска документов, сформированный по формуле (6), имеет следующий вид:

$$Q_1^{t+} = \bigcup_l (s_l : s_l \in Q^s) = s_i \cup s_j \cup ... \cup s_k =$$

$$= (t : t \in s_i) \cup (t : t \in s_j) \cup ... \cup (t : t \in s_k).$$

Поисковая выборка документов из коллекции D по запросу Q_1^{t+} формируется как пересечение индекса документов и запроса:

$$Q_1^{t+} \cap T^D = ((t : t \in s_i) \cup (t : t \in s_j) \cup \; ... \; \cup (t : t \in s_k)) \cap T^D =$$

$$= ((t : t \in s_i) \cap T^D) \cup ((t : t \in s_j) \cap T^D) \cup ... \cup ((t : t \in s_k) \cap T^D) .$$

Поскольку пользователь при задании поискового запроса под термином t имел в виду одно конкретное значение, положим, что это значение фиксируется за синсетом s_i. Тогда согласно процедуре разрешения омонимии $Q_s = \{s_i\}$, а модифицированный поисковый запрос имеет вид

$$Q_2^{t+} = \bigcup_l (s_l : s_l \in Q^s) = s_i = (t : t \in s_i) .$$

Соответственно, выбор документов происходит по формуле

$$Q_2^{t+} \cap T^D = (t : t \in s_i) \cap T^D .$$

Согласно [13] точность поиска определяется по формуле

$$P = \frac{|D_{\text{п.рел}}|}{|D_{\text{п.общ}}|},$$

где $D_{\text{п.общ}}$ — общее множество полученных документов.

Каждое выражение $(t : t \in s_l) \cap T^D$, где $l \in \{i, j, ..., k\}$, определяет выборку документов D_l. Пусть $D_l^{\text{рел}} \subseteq D_l$ — множество документов, соответст-

вующих поисковым интересам пользователя. Тогда точность поиска для случая Q_1^{t+} можно рассчитать по формуле

$$P_1 = \frac{\left|D_i^{\text{рел}} \cup D_j^{\text{рел}} \cup ... \cup D_k^{\text{рел}}\right|}{\left|D_i \cup D_j \cup ... \cup D_k\right|}.$$

Поскольку в запросе только значение синсета s_i отвечает поисковым интересам пользователя, получим

$$\left|D_i^{\text{рел}}\right| \geq 0, \text{ а } \left|D_j^{\text{рел}}\right| = ... = D_k^{\text{рел}} = 0.$$

Отсюда следует

$$P_1 = \frac{\left|D_i^{\text{рел}} \cup D_j^{\text{рел}} \cup ... \cup D_k^{\text{рел}}\right|}{\left|D_i \cup D_j \cup ... \cup D_k\right|} = \frac{\left|D_i^{\text{рел}}\right|}{\left|D_i \cup D_j \cup ... \cup D_k\right|}.$$

Для случая Q_2^{t+}

$$P_2 = \frac{\left|D_i^{\text{рел}}\right|}{\left|D_i\right|}.$$

Очевидно, что $P_2 \geq P_1$. Таким образом, без разрешения омонимии при отображении $Q^t \to Q^s$ точность поиска уменьшается, что и требовалось доказать.

Согласно теореме 2 при наличии в начальном запросе многозначных терминов обязательно должна быть разрешена проблема омонимии [7]. При этом пользователем фиксируется определенный синсет $s \in \{s_i, s_j, ..., s_k\}$, наиболее соответствующий поисковым интересам пользователя. Решение о фиксации определенного значения s принимается на основании описательной части синсета D из (1), которая предлагается пользователю интеллектуальным редактором запроса [8]. Таким образом, на выходе процедуры отображения $\varphi(t)$ получаем однозначное преобразование Q^t в Q^s.

Формализация процедуры вертикальной коррекции поискового запроса. В основе онтологии предметной области, подобно WordNet [9], лежит упорядоченная иерархия синсетов. Ее базовую структуру можно рассматривать как дерево, в корне которого находится наиболее общий синсет, выражающий наиболее абстрактное понятие для данной онтологии, а на каждом следующем уровне синсеты конкретизируются. Таким образом, терминальные узлы дерева являются наименее общими и описывают соответственно очень узкие понятия предметной области.

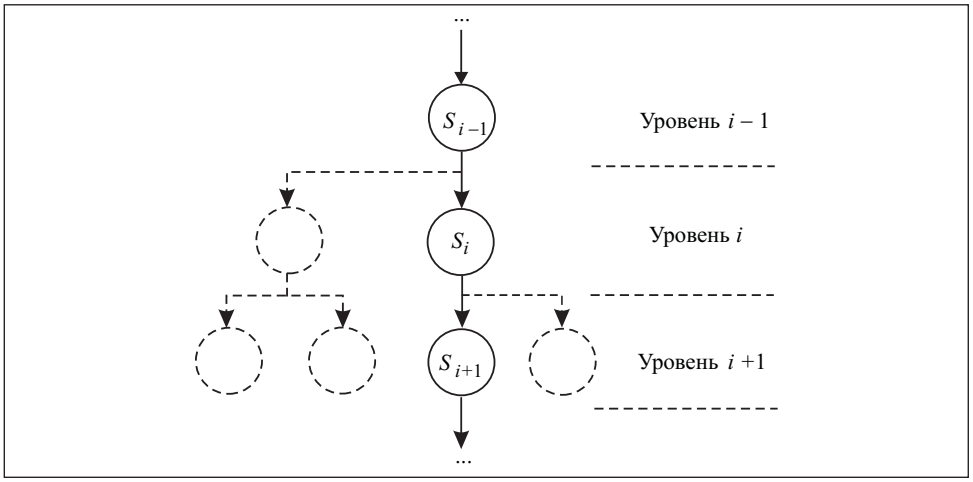

Рис. 1. Фрагмент иерархии онтологии

С учетом этой особенности структуры онтологии в [7] предложено для модификации начального поискового запроса использовать специальный модуль редактора запроса, предназначенный для навигации по иерархии понятий онтологии. Этот модуль позволяет провести визуализацию части дерева синсетов и, используя иерархичность отношений между ними, дать возможность пользователю изменять уровень абстракции запроса, передвигаясь по иерархии вниз или вверх, что непосредственно влияет на результирующую поисковую выборку документов.

В данной процедуре квазисемантического поиска используется отношение $r_h \in R$ онтологии для модификации поискового запроса Q^s. Рассмотрим фрагмент иерархии онтологии, приведенный на рис. 1.

Пусть для наглядности номер синсета соответствует уровню общности или абстрактности понятия. Чем ниже уровень, тем более общим (абстрактным) является понятие. Поскольку на каждой итерации процедуры модификации поискового запроса механизм навигации может быть использован только для одного из понятий начального поискового запроса, после выполнения преобразования (5) получим $Q^s = \{s_i\}$.

Согласно правилам выполнения навигации по иерархии модификация запроса происходит в результате перехода от понятия s_i к более общему понятию s_{i-1} или к более конкретному s_{i+1}. При этом запрос Q^s может быть соответственно модифицирован:

$$Q^{s+} = (Q^s \setminus \{s_i\}) \cup \{s_{i-1}\}, \tag{8}$$

$$Q^{s+} = (Q^s \setminus \{s_i\}) \cup \{s_{i+1}\}. \tag{9}$$

В (8) и (9) подразумевается модификация поискового запроса с помощью синсета (s_{i-1} или s_{i+1}) посредством замены текущего синсета s_i в запросе новым (соответственно s_{i-1} или s_{i+1}). Следовательно, выбор любой из стратегий приведет к вхождению нового синсета (s_{i-1} или s_{i+1}) в состав поискового запроса в зависимости от того, куда будет двигаться во время навигации по дереву онтологии пользователь — в направлении гиперонимизации (генерализации) или в направлении гипонимизации (конкретизации).

Согласно (3) и (4) возможны два варианта поискового индекса документов: классический полнотекстовой индекс и концепт-индекс, когда документы индексируются не терминами, а синсетами. Поэтому необходимо рассмотреть механизм работы квазисемантической процедуры вертикальной модификации для каждого из этих случаев.

В случае, когда документы индексируются согласно (3), процесс поиска документов происходит на основе терминов. Гипонимизация синсета $s_i \in Q^s$ с использованием синсета s_{i+1}, согласно условиям формирования иерархических связей в онтологии, конкретизирует термины $t \in s_i$ из запроса терминами $t \in s_{i+1}$. При этом пересечение множеств индексов документов T^D и терминов сенсета s_{i+1} в составе модифицированного поискового запроса будет давать гипонимизированную, с семантически более узкими рамками, выборку документов, нежели пересечение того же множества T^D и терминов синсета s_i.

Следовательно, поскольку полнотекстовой индекс документов в общем случае строится без учета связей онтологии, иерархическое отношение между терминами синсетов прямо не отображается на таком индексе. Если термин гипонима более употребляем в контексте данной коллекции документов, то, в соответствии с правилами функционирования полнотекстовой поисковой системы, количество документов по такому модифицированному запросу будет больше, чем по немодифицированному. Однако такая выборка в большей мере отвечает поисковым интересам пользователя.

Таким образом, гипонимизированная поисковая выборка характеризуется повышенным уровнем пертинентности по отношению к поисковым потребностям пользователя. Из этого вытекает и обратное свойство: если провести модификацию поискового запроса посредством гиперонимизации синсета, поднимаясь по иерархии онтологии вверх, то пересечение множества T^D и терминов синсета s_{i-1} будет давать соответственно гиперонимизированную поисковую выборку документов, которая включает документы, соответствующие более общей интерпретации начального немодифицированного поискового запроса.

Если рассматривать вариант поисковой системы на основе концепт-индекса, когда индексация документов проводится с помощью синсетов

онтологии [11] согласно (4), то и процедура поиска документов в таком случае будет происходить на основе синсетов без перехода на уровень терминов. Поскольку процедура создания такого индекса с самого начала тесно связана с использованием онтологии предметной области, семантика связей между синсетами может непосредственно влиять на результаты поиска. Так, отношение гипоним/гипероним может быть использовано при формировании результатов поиска, когда в состав полной выборки по синсету, согласно требованиям пользователя, включена выборка по синсету, объединенная с выборкой по его дочерним элементам.

Выборка по дочерним элементам может иметь определенные ограничения по глубине. Ранг результатов, выбираемым по дочерним синсетам, может быть уменьшен при использовании соответствующего коэффициента, зависящего от глубины (числа переходов) соответствующей сложенной иерархической связи. Кроме того, присоединение результатов по дочерним синсетам, вероятно, целесообразно выполнять только на достаточно глубоких ветках иерархии, где иерархическая связь проявляется более значительно. В данном случае конкретное решение следует принимать при проектировании реальной поисковой системы.

Теорема 3. При выполнении поиска результатов в концепт-индексе модификация квазисемантического поискового запроса посредством гипонимизации синсета согласно выражению $Q^{s+} = (Q^s \setminus \{s_i\}) \cup \{s_{i+1}\}$ обеспечивает выборку документов более конкретного характера, а модификация посредством его гиперонимизации согласно выражению $Q^{s+} = (Q^s \setminus \{s_i\}) \cup \{s_{i-1}\}$ обеспечивает более абстрактный характер выборки.

Доказательство. Пусть D_i и D_{i-1}, D_{i+1} — выборки документов соответственно только по синсету s_i и только по синсетам s_{i-1}, s_{i+1}, а $D(s)$ — множество полной выборки документов по некоторому синсету s с учетом автоматического включения в запрос соответствующих поддеревьев. Тогда

$$D(s_{i+1}) = D_{i+1},$$

$$D(s_i) = D_i \cup D(s_{i+1}) = D_i \cup D_{i+1},$$

$$D(s_{i-1}) = D_{i-1} \cup D(s_i) = D_{i-1} \cup D_i \cup D_{i+1}.$$

Таким образом, модификация текущего запроса синсетом s_{i+1} (т.е. замена синсета из текущего запроса его гипонимом) автоматически отсекает от результирующей выборки ту часть документов D_i, которая соответствовала только синсету s_i, увеличивая концентрацию документов, пертинентных поисковому интересу пользователя. И наоборот, модификация запроса синсетом s_{i-1} (т.е. использование гиперонима синсета из

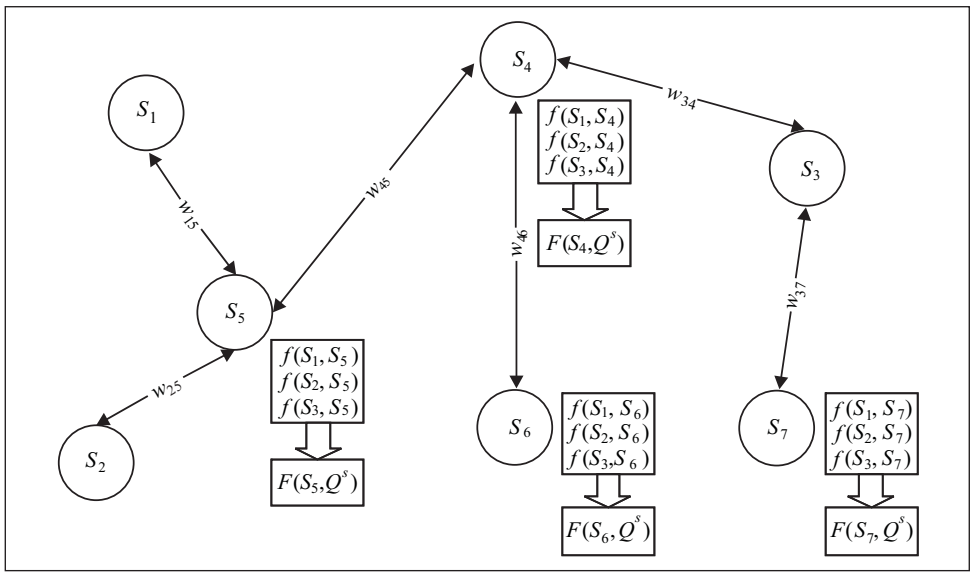

Рис. 2. Схема процедуры поиска семантического среза онтологии

текущего запроса) обеспечивает расширение результирующей выборки за счет документов D_i, содержание которых может потенциально отвечать поисковым интересам пользователя, но в более абстрактном или общем виде вследствие гиперонима синсета s_i, что и требовалось доказать.

Согласно теореме 3 механизм вертикальной модификации поискового запроса посредством навигации по иерархическим связям онтологии, при поддержке редактора запроса, дает возможность управлять качественными (а иногда и количественными) характеристиками поисковой выборки по модифицированному поисковому запросу. При высоком уровне разрозненности результатов поиска или его большом объеме пользователь имеет возможность сузить поле поиска и соответственно выделить пертинентные документы. Когда коллекция документов не дает ожидаемого отклика (небольшое количество результатов), у пользователя есть возможность ослабить условия поиска, расширив итоговое множество документов, соответствующих поисковому запросу.

Формализация процедуры ассоциативной коррекции поискового запроса. Основная цель этой процедуры квазисемантического поиска — предложить пользователю возможные варианты нелинейной модификации поискового запроса на основе ассоциативных связей между синсетами. В этом случае онтология может рассматриваться как семантическая сеть понятий (рис. 2). Выделенные из текущего поискового запроса понятия фиксируются в этой сети, и выполняется процедура поиска других

синсетов онтологии, которые семантически близки всем или части понятий из запроса, т.е. определяется их семантическое окружение или семантический срез онтологии.

Варианты модификации — это взвешенное множество синсетов онтологии, которые с определенной пропорциональностью семантически близки множеству синсетов из текущего поискового запроса. При этом запрос рассматривается как отображение терминов запроса во множестве синсетов Q^s. Пусть на рис. 2 это синсеты $\{s_1, s_2, s_3\}$, а семантическая близость определяется по определенному алгоритму на основе семантических связей в онтологии.

Нелинейность модификации поискового запроса достигается вследствие того, что семантический срез может представлять собой области онтологии, напрямую не связанные с понятием из начального запроса, но в результате активизации ассоциативных связей между синсетами могут выявлять скрытые пути модификации поискового запроса в рамках поисковых интересов пользователя.

Пусть $Q^s = \{s_1, ..., s_n\}$, где n — число синсетов в поисковом запросе. Для решения задачи поиска среза онтология рассматривается как семантическая сеть, поэтому учитываются все связи r_h и r_a. Каждый тип связи может иметь собственный вес: соответственно w_h и w_a. Более сложный и обобщенный вариант ассоциативной коррекции возможен в случае, когда каждое конкретное отношение между двумя синсетами (s_i и s_j) онтологии в рамках определенного типа отношений определяется собственным весом w_{ij} (см. рис. 2). Значение этого веса зависит от силы семантической связи между соответствующими понятиями в онтологии.

Процедуру нахождения семантического среза можно рассматривать как последовательность трех этапов (рис. 3).

1. Для каждого $s_i \in Q^s$ и каждого $s_j \in S$, где $i \neq j$, рассчитывается функция семантической близости синсетов $f(s_i, s_j)$. Известен ряд подходов к решению этой задачи. Так, в работе [15] рассмотрено пять различных методик поиска семантической близости двух понятий на основе связей онтологии. Для квазисемантического поиска возможно использование только одной из них, а именно:

$$f(s_i, s_j) = C - path(s_i, s_j) - dk, \qquad (10)$$

где C — наибольший возможный радиус семантического влияния узла s_i; $path(s_i, s_j)$ — самый короткий путь от s_i до s_j; d — число смен типов семантического отношения (так называемых поворотов) на пути от s_i к s_j; k — некоторая константа, определяемая экспериментально.

Соотношение (10) признано описанием «самой медленной» из пяти рассмотренных в [15] методик, однако необходимо заметить, что только

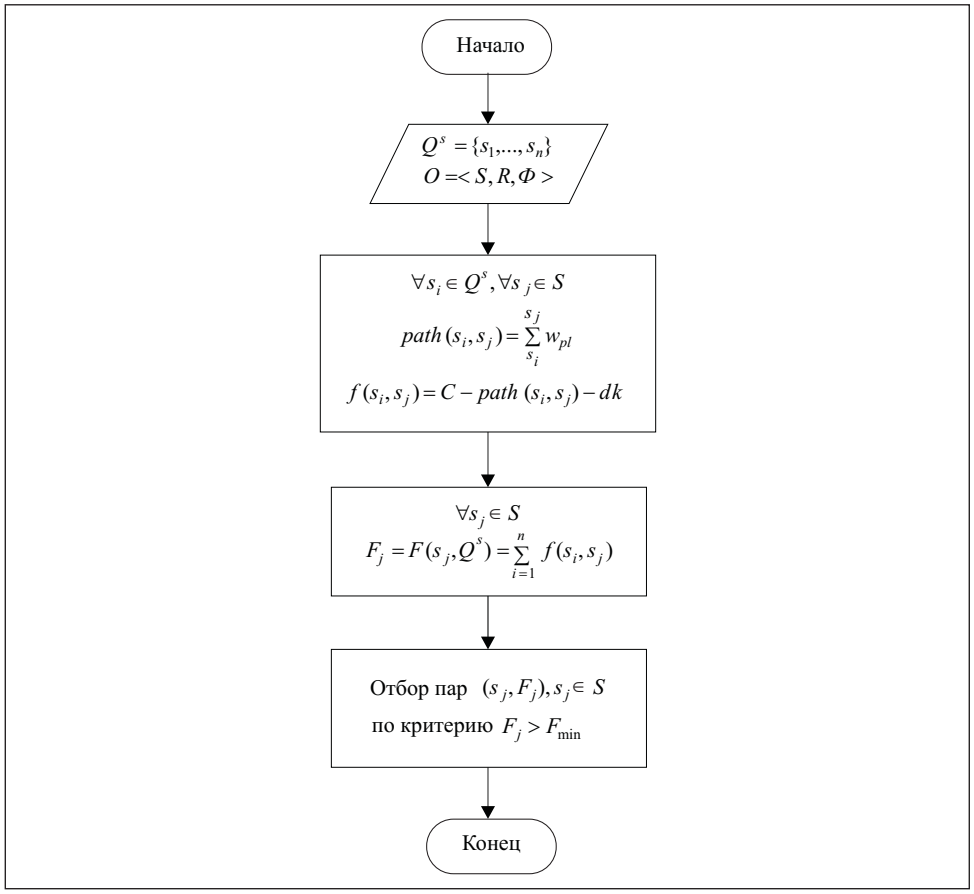

Рис. 3. Блок-схема алгоритма процедуры поиска семантического среза

данная функция позволяет использовать все типы связей в онтологии. Высокое быстродействие других методик достигается только при использовании отношения гипонимии, что неприемлемо в связи с поставленной задачей поиска среза с учетом всех типов связей онтологии.

Длина пути рассчитывается по обобщенной формуле

$$path(s_i, s_j) = \sum_{s_i}^{s_j} w_{pl},$$

где w_{pl} — вес отношения между стоящими рядом синсетами s_p и s_l, которые являются промежуточными узлами на самом коротком пути между синсетами s_i и s_j.

Процедура модификации запроса	Характеристика поисковой выборки	
	Полнота	Точность
Проекция начального запроса на онтологию и горизонтальная коррекция поискового запроса	Расширение результирующей поисковой выборки достигается посредством добавления в начальный поисковый запрос синонимических рядов синсетов онтологии. Пример: начальный запрос: база данных модифицированный запрос в терминах: база данных, бд, хранилище данных	Разрешение проблемы омонимии на основе онтологии и при непосредственном участии пользователя позволяет исключить присутствие непертинентных объектов в поисковой выборке документов. Пример: начальный запрос: экология модифицированный запрос: экология (раздел биологии) или экология (состояние окружающей среды)
Вертикальная коррекция поискового запроса	Модификация понятием, стоящим выше по иерархии онтологии, чем текущее понятие запроса, позволяет включать в поисковую выборку более широкий по уровню обобщения круг результатов, соответствующих более обобщенным (в противовес конкретным) терминам. Это дает возможность повышать содержательную полноту результатов добавлением обобщенных (абстрактных) документов по отношению к начальному поисковому запросу. Пример: начальный запрос: естественная наука модифицированный запрос: наука	Модификация поискового запроса гипонимом к синсету из текущего поискового запроса делает результирующую выборку семантически более узкой, т.е. способной повысить пертинентность результатов поиска в соответствии с поисковыми намерениями пользователя, или такой, которая включает лишь некоторый подкласс результатов из перечня документов по более абстрактному (начальному) поисковому запросу. Пример: начальный запрос: естественная наука модифицированный запрос: география
Ассоциативная коррекция поискового запроса	Поскольку такая процедура обеспечивает возможность нелинейной модификации поискового запроса, расширение содержательной полноты результатов поиска достигается добавлением потенциально семантически связанных документов. Семантическая связанность достигается с помощью семантических связей между понятиями онтологии, которыми модифицируется поисковый запрос. Пример: начальный запрос: лингвистика, база знаний модифицированный запрос: лингвистика, база знаний, онтология	Повышается пертинентность результатов поиска, так как поиск семантических срезов позволяет найти скрытые латентные семантические связи и предложить пользователю понятия, которые могут помочь охарактеризовать более точно его поисковый интерес. Пример: начальный запрос: лингвистика, программное обеспечение модифицированный запрос: лингвистика, программное обеспечение, программа-переводчик

2. Для каждого $s_j \in S$ рассчитывается интегральная величина F_j близости синсета s_j к поисковому запросу Q^s:

$$F_j = F(s_j, Q^s) = \sum_{i=1}^{n} f(s_i, s_j).$$

3. Выполняется сортировка всех синсетов онтологии $s_j \in S$ по значению их интегральной величины F_j. Из полученной таблицы пар (s_j, F_j) отбираются пары, для которых $F_j > F_{\min}$, где $F_{\min}$ — минимальный порог. Результатом этого этапа есть множество пар (s, F), семантически наиболее близких к поисковому запросу Q^s, при этом F определяет меру семантической близости соответствующего синсета s.

Редактор запроса использует найденное множество пар (s, F) для визуализации и предложения пользователю семантического среза онтологии по текущему поисковому запросу. Величина F для каждого отобранного синсета s указывает на вес семантической близости к запросу Q^s и позволяет выполнить визуальное ранжирование выбранных понятий. На основе этих данных пользователь, при необходимости опираясь на описание понятий, принимает решение относительно модификации запроса синсетами из предложенного семантического среза, что позволит в случае необходимости соответствующим образом изменить направление поиска и перейти в ту область результатов, которая больше отвечает поисковым интересам пользователя.

Исследование влияния процедур модификации запроса на характеристики квазисемантической поисковой выборки. Рассмотренные три процедуры направлены на формирование оптимизированного запроса на квазисемантический поиск и позволяют априори итерационно управлять поисковым откликом. Можно выделить две основные характеристики поискового результата, на которые упомянутые процедуры влияют: полнота и точность результатов поиска.

В таблице приведены данные о влиянии каждой рассмотренной процедуры модификации квазисемантического поискового запроса на указанные характеристики.

В случае ассоциативной коррекции поискового запроса анализировать точность поисковой выборки в классическом ее понимании нельзя. Основная задача этой процедуры — нелинейный переход в область документов, семантически связанных (через онтологию) с поисковой выборкой по немодифицированному поисковому запросу. Такой переход способен обеспечить точность результатов, удовлетворяющую ожидания пользова-

теля. Именно поэтому в данном случае повышение точности можно рассматривать как повышение пертинентности.

В тоже время, полнота поисковой выборки в этом случае также не является классической полнотой, поскольку она достигается вследствие охвата других областей онтологии во время формирования поискового запроса, что можно рассматривать как скрытую содержательную полноту. Поиск по модифицированному запросу, который сформирован согласно описанному принципу, позволяет включить релевантные документы из соседних областей предметной области.

Выводы

Таким образом, использование предложенных способов реализации квазисемантического поиска при взаимодействии с полнотекстовым индексом и концепт-индексом позволяет целенаправленно повышать пертинентность результатов поиска. Включение в поисковую выборку семантически связанных документов дает возможность дополнительно расширить поле потенциально пертинентных документов, что достигается при реализации трех стратегий коррекции поискового запроса: горизонтальной (синонимических связей онтологии), вертикальной (иерархических связей онтологии) и ассоциативной (оригинального алгоритма поиска семантического среза онтологии на основе текущего поискового запроса).

Рассмотренная специфика отображения поискового запроса из поля терминов в поле синсетов онтологии позволяет решать проблему омонимии интерпретации поискового запроса на этапе его формирования. Активную роль в этом процессе играет интерактивное взаимодействие с пользователем, который является основным источником поискового интереса.

Formal models of the procedures to involve a semantic component into search query used in quasi-semantic search are proposed. The query is modified using a linguistic ontology of the given subject field. The specificity of the search query mapping from the field of terms to the ontology synsets field is considered. The three search query correction strategies (horizontal, vertical and associative) are proposed and investigated. The impact of the search query modification procedures on the search query results characteristics is analyzed.

1. *Вдовіченко А. В.* Інтелектуалізовані пошукові системи. Класифікація та порівняння // Штучний інтелект. — 2002. — № 3. — С. 61—70.
2. *Ландэ Д. В.* Поиск знаний в Internet. Профессиональная работа. — М. : Изд. дом «Вильямс», 2005. — 272 с.
3. *Ландэ Д. В., Снарский А. А., Безсуднов И. В.* Интернетика. Навигация в сложных сетях: модели и алгоритмы. — М. : Книжный дом «ЛИБРОКОМ», 2009. — 264 с.
4. *Додонов О. Г., Ланде Д. В., Путятін В. Г.* Інформаційні потоки в глобальних комп'ютерних мережах. — К. : Наук. думка, 2009. — 294 с.

5. *Згуровський М. З., Якименко Ю. І., Тимофєєв В. І.* Інформаційні мережеві технології в науці і освіті // Системні дослідження та інформаційні технології. — 2002. — № 3. — С. 43—56.
6. *Марченко О. О.* Алгоритми семантичного аналізу природомовних текстів : Автореф. дис. … канд. фіз.-мат. наук. — Київ, 2005. — 17 с.
7. *Михайлюк А. Ю., Пилипчук О. В., Сніжко М. В., Тарасенко В. П.* Квазісемантичний пошук текстових даних в електронному інформаційному ресурсі // Радіоелектроніка та інформатика. — 2009. — № 3. — С. 61—67.
8. *Михайлюк А. Ю., Пилипчук О. В., Сніжко М. В., Тарасенко В. П.* Спосіб структурно-алгоритмічної організації інтелектуального редактора запиту на квазісемантичний пошук//Міжнародна науково-технічна конф. «Інтелектуальні технології лінгвістичного аналізу». Тези доповідей. — К. : Вид-во «НАУ-друк», 2009. — С. 43.
9. *Соловьев В. Д., Добров Б. В., Иванов В. В., Лукашевич Н. В.* Онтологии и тезаурусы: Учеб. пособие.— Казань : Казанский госуниверситет, 2006. — 198 с.
10. *Clarke C., Cormack G.* Dynamic Inverted Indexes for a Distributed Full-Text Retrieval System. TechRep MT-95-0. University of Waterloo. — 1995. — 13 p.
11. *Добров Б. В., Лукашевич Н. В.* Тезаурус и автоматическое концептуальное индексирование в университетской информационной системе РОССИЯ // Третья Всероссийская конференция по электронным библиотекам «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». — Петрозаводск: КарНЦ РАН, 2001 — С.78—82.
12. *Гаврилова Т. А., Хорошевский В. Ф.* Базы знаний интеллектуальных систем. — СПб : Питер, 2000. — 384 с.
13. *Дорнфест Р., Бош П., Калишейн Т.* Секреты Google. Трюки и тонкая настройка. — М. : «Русская редакция», 2008. — 512 с.
14. *Manning C., Prabhakar R., Schutze H.* An Introduction to Information Retrieval.— Cambridge: Cambridge University Press, 2009. — 581 p.
15. *Budanitsky A., Hirst G.* Evaluating WordNet-based measures of lexical semantic relatedness // Computational Linguistics. — 2006. — Vol.32, № 1. — P. 13—47.

Поступила 01.12.10;
после доработки 07.02.11

ЗАМЯТИН Денис Станиславович, канд. техн. наук, доцент кафедры специализированных компьютерных систем Национального технического университета Украины «Киевский политехнический ин-т», который окончил в 2000 г. Область научных исследований — методы и средства поиска и логической систематизации распределенных данных.

МИХАЙЛЮК Антон Юрьевич, канд. техн. наук, зав. научно-исследовательской лабораторией информатизации образования Киевского университета им. Б. Гринченко. В 1985 г. окончил Киевский политехнический ин-т. Область научных исследований — методы и средства интеллектуального анализа данных.

МИХАЙЛЮК Елена Станиславовна, науч. сотр. кафедры специализированных компьютерных систем Национального технического университета Украины «Киевский политехнический ин-т», который окончила в 1989 г. Область научных исследований — экспертные системы, их применение в задачах анализа данных.

ПЕТРАШЕНКО Андрей Васильевич, канд. техн. наук, доцент кафедры специализированных компьютерных систем Национального технического университета Украины «Киевский политехнический ин-т», который окончил в 2000 г. Область научных исследований — методы поиска информации в несистематизированном гетерогенном ресурсе.

ПИЛИПЧУК Алексей Васильевич, аспирант кафедры специализированных компьютерных систем Национального технического университета Украины «Киевский политехнический ин-т», который окончил в 2008 г. Область научных исследований — поиск информации, квазисемантический поиск текстовых данных.

ТАРАСЕНКО Владимир Петрович, д-р техн. наук, профессор, зав. кафедрой специализированных компьютерных систем Национального технического университета Украины «Киевский политехнический ин-т», который окончил в 1968 г. Область научных исследований — повышение эффективности обработки ресурсов глобального электронного информационного пространства.