ФОРМИРОВАНИЕ МНОЖЕСТВА СВЯЗНЫХ КОНЦЕПТОВ ДЛЯ АВТОМАТИЧЕСКОГО СИНТЕЗА ОНТОЛОГИЙ

Лариса Чалая, Антон Чижевский

***Аннотация**: В данной работе предлагается метод нахождения концептов с учетом связей между ними, позволяющий повысить качество автоматически создаваемых онтологий. Метод может эффективно использоваться для задач семантического поиска в системах анализа электронных текстов и автоматического создания онтологических моделей предметной области.*

***Ключевые слова**: связные концепты, онтология, морфологические свойства слов, автоматический синтез онтологий, предметная область*

***ACM Classification Keywords**: I.2 Artificial Intelligence - I.2.4 Knowledge Representation Formalisms and Methods*

Введение

В настоящее время отмечается важность использования онтологии как основы для спецификации и разработки программного обеспечения, поддержки общего доступа к информации, поиска информации, взаимодействия при объединении информации, создании порталов знаний, разработке пользовательского интерфейса программных систем, редакторов информации и интеллектуальных систем [Артемьева, 2008]. Интернет уже давно стал огромным хранилищем полезной информации, ориентированным на людей. Однако чем больше такой информации в Интернете, тем труднее найти в этом хранилище релевантную информацию. Проект семантического Интернета направлен на решение этой проблемы путем хранения в Интернете онтологий, в терминах которых представлена информация, и благодаря унификации ее представления. Однако и здесь возникла похожая проблема — кто и как будет создавать и сопровождать онтологии и обеспечивать унифицированное представление информации в Интернете [Артемьева, 2006]. В связи с этим в последнее время получили широкое развитие исследования в области автоматического синтеза онтологических моделей, позволяющих повысить эффективность систем семантического поиска по запросам пользователей (в корпусе текстов, электронных библиотеках, в сети Интернет) [Чалая, 2012]. Качество формируемых онтологий, используемых для создания поисковых систем, во многом определяется полнотой учета в онтологической модели наиболее значимых концептов для корпуса анализируемых текстов с учетом их тематической специфики (под концептами будем в дальнейшем понимать наиболее значимые слова и словосочетания в анализируемом тексте, которые могут быть учтены в онтологической модели). В связи с этим целесообразно решить задачу формирования множества концептов будущей онтологии с учетом связей между ними.

Постановка задачи

Методы нахождения концептов при автоматическом синтезе онтологий и нахождения шаблонных связей между ними (типа «часть-целое» и «отношение») рассматриваются, в частности, в работах [Чалая, 2012; Чалая, 2011; Зябрев, 2010]. Однако результаты экспериментального исследования этих методов показали, что при поиске слов и словосочетаний, которые могут использоваться в качестве концептов,

сформированное множество концептов-претендентов не всегда соответствует такому же множеству, составленному экспертом предметной области. Это приводит к тому, что некоторые важные понятия предметной области могут не попасть в автоматически создаваемую онтологическую модель. Кроме того, в этих методах отсутствует процедура общего ранжирования по значимости списка всех концептов-претендентов, а осуществляется лишь раздельное ранжирование слов и словосочетаний, входящих в этот список.

Целью данной статьи является модификация и программная реализация методов нахождения концептов с учетом связей между ними для повышения качества автоматически создаваемых онтологий. При этом особое внимание уделяется задачам ранжирования однословных/многословных концептов и выявления отношений между ними.

Поиск концептов для онтологии

Для решения задачи поиска концептов будущей онтологии целесообразно использовать наиболее перспективные алгоритмы ранжирования документов по релевантности запросу. В качестве базового алгоритма рассмотрим алгоритм atr, приведенный в работе [Гулин, 2006]. Здесь для каждого запроса вычисляется суммарная характеристика *Score* (показатель релевантности документа запросу), определяемая следующим образом:

$$Score = W_{single} + W_{pair} + k_1 * W_{AllWords} + k_2 * W_{Phrase} + k_3 * W_{HalfPhrase} + W_{PRF} \quad (1)$$

где: W_{single} – частная характеристика, отражающая встречаемость слов из запроса в документе;

W_{pair} – частная характеристика, отражающая встречаемость пар слов из запроса в документе;

$W_{AllWords}$ – частная характеристика, отражающая наличие всех слов запроса в документе;

W_{Phrase} – частная характеристика, отражающая встречаемость полного текста запроса в документе;

$W_{HalfPhrase}$ – частная характеристика, отражающая наличие нескольких слов запроса в одном предложении документа;

W_{PRF} – частная характеристика, основанная на добавлении в запрос дополнительных релевантных слов (подход «Pseudo-relevance feedback» [Jinxi Xu, 1996]);

k_1, k_2, k_3 – весовые коэффициенты.

Идея базового алгоритма состоит в формировании списка релевантных текстов для претендента на концепт из корпуса текстов, на основе которого проектируется онтология. Чем больше будет найдено релевантных документов по запросу, тем более важным для рассматриваемой предметной области является слово (словосочетание), представляющее данный запрос, и, следовательно, оно может с большей вероятностью претендовать на включение в ранжированный список концептов-претендентов.

Отметим, что для составления ранжированного списка претендентов на концепты будущей онтологии не обязательно проводить поиск по всем отдельным словам. Целесообразнее ранжировать лишь слова, найденные как важные для предметной области [Чалая, 2012]. Кроме того, необходимо провести ранжирование словосочетаний в тексте для последующего определения концептов-претендентов. При этом следует учесть максимум словосочетаний, выделенных в тексте, отбросив лишь однозначно не подходящие на роль концептов будущей онтологии.

По морфологическим свойствам главного слова словосочетания классифицируются следующим образом: глагольные (составить план, стоять у доски); именные (план сочинения, поездка по городу); наречные (крайне важно, вдали от дороги). В разрабатываемом методе будут рассматриваться только именные словосочетания, поскольку ни глагольные, ни наречные словосочетания не могут претендовать на роль концептов онтологии.

Для определения типов выделенных словосочетаний и количественного оценивания силы связи между словоформами в рамках исследуемого текста используем алгоритм Гинзбурга [Гинзбург, 1998]. Этот алгоритм предназначен для поиска контекста слов в рамках рассматриваемого текста и предполагает последовательную реализацию следующих этапов:

1. Определение в предложенном тексте T относительной частоты $RFT(x)$ для каждой словоформы x;
2. Нахождение в тексте T предложения T_W, которые содержат слово W. Из совокупности предложений T_W строится частотный словарь $V(T_W)$, который содержит относительную и абсолютную частоты. Относительную частоту словоформы x в $V(T_W)$ обозначим $RFT_W(x)$;
3. Затем сравниваются относительные частоты T и T_W и вводится индекс значимости $SI(x)$ словоформы x в контексте слова W. Он вычисляется по формуле:

$$SI(x) = \frac{RFT_W(x)}{RFT(x)}, \tag{2}$$

если индекс значимости больше единицы, словоформу x можем относить к контексту слова W.

После определения контекстных множеств появляется возможность вычислить силу зависимости между двумя словоформами. В частности, для вычисления силы связи (BF) между словоформами x_1 и x_2 необходимо составить контекстные множества S_1 и S_2, которые имеют положительный индекс значимости. Число словоформ в S_1 обозначим N_1, в S_2 – N_2. Множества могут иметь две части: общую для S_1 и S_2 и специфичную для S_1 или S_2.

Алгоритм обработки множеств S_1 и S_2 [Воронина, 2010]:

1) Специфические части суммируются по модулю (SS);
2) Для каждой словоформы в общей части вычисляется разность: $SI(S_1) - SI(S_2)$;
3) Полученные разности суммируются по модулю (SG);
4) Вычисляется общая сумма $(ST : ST = SS + SG)$;
5) Рассчитывается сумма всех индексов значимости всех словоформ, которые принадлежат множеству S_1 $(SumS_1)$;
6) Рассчитывается сумма всех индексов значимости всех словоформ, которые принадлежат множеству S_2 $(SumS_2)$;
7) Определяется сила связи между словоформами x_1 и x_2:

$$BF = 1 - \frac{ST}{SumS_1 + SumS_2} \tag{3}$$

Полученный результат находится в диапазоне от 0 до 1. При этом 0 соответствует полному отсутствию общих слов в контексте, а 1 – полному совпадению множеств слов контекста и их частот. Величину силы

связи между словоформами также можно интерпретировать и как величину, обратную расстоянию между словоформами в семантическом пространстве исследуемого текста. Для каждого слова рассматриваемого словосочетания необходимо составить список слов, связанных с ним. В этот список заносят слова, у которых сила такой связи не меньше, чем $minBF$:

$$minBF = K_1 * max(BF), \tag{4}$$

где K_1 – коэффициент, задающий порог для выбора из множества возможных контекстов лишь тех слов, у которых сила связи с рассматриваемым словом является существенной (рекомендуемый диапазон значений K_1: 0.3-0.5).

Исследования показали, что в глагольных, наречных, количественных словосочетаниях, а также в словосочетаниях, у которых главным словом является местоимение, сила связи между словами всегда меньше показателя $minBF$. Кроме того, этот показатель, как правило, меньше у именных словосочетаний, состоящих из существительных, но не несущих семантическую нагрузку (например, «обращение к первоисточнику», «работа в команде»). Предварительное удаление таких словосочетаний из исходного анализируемого множества позволяет существенно ограничить количество возможных концептов проектируемой онтологии.

Для последующего ранжирования отобранных словосочетаний используем модифицированную формулу (1), из которой предлагается убрать слагаемое W_{single} . Целесообразность такой модификации обусловлена тем, что запросы, которые обрабатываются с помощью исходной формулы, вводятся не пользователем, а представляют собой словосочетания, найденные в тексте. Если такие словосочетания претендуют на роль концептов онтологии, то они должны быть близкими к смыслу текста в исходном виде, а, следовательно, встречаемость в тексте отдельных слов из этих словосочетаний не должна влиять на суммарную характеристику $Score$. Кроме того, предлагается модифицировать и слагаемое W_{PRF} . Модифицированный подход «Pseudo-relevance feedback» предполагает реализацию двухэтапной процедуры поиска коцептов. На первом этапе мы используется базовый метод определения W_{PRF} , описанный в [Jinxi Xu, 1996]. На втором этапе для документов, соответствующих первым позициям формируемого списка и отнесенных к классу релевантных, определяется дополнительная совокупность документов, в которых содержатся ссылки на другие документы коллекции (например, слагаемое W_{PRF} может непосредственно определяться количеством ссылок на уже найденные релевантные документы). Этот подход наиболее эффективен при создании онтологии на основе корпуса научных текстов: если в одном таком тексте содержатся ссылки на другой текст корпуса, то высока вероятность того, что они имеют близкую тематику.

Модифицированная формула для определения суммарной характеристики релевантности текстов для найденных словосочетаний ($Score_1$) принимает следующий вид:

$$Score_1 = W_{pair} + k_1 * W_{AllWords} + k_2 * W_{Phrase} + k_3 * W_{HalfPhrase} + W_{PRF} \tag{5}$$

После того как будут найдены списки релевантных документов по всем найденным словосочетаниям, необходимо определить целесообразность отнесения наиболее значимых словосочетаний к классу концептов проектируемой онтологии.

Для этого вводится понятие $minQ$:

$$minQ = K_2 * Q, \tag{6}$$

где K_2 – пороговый коэффициент, позволяющий удалить из множества найденных словосочетаний такие, для которых найдено слишком мало релевантных текстов в корпусе; Q – общее количество текстов в корпусе.

Таким образом, к концептам будущей онтологии могут быть отнесены лишь те словосочетания, для которых количество релевантных текстов в корпусе не меньше, чем $minQ$. Коэффициент K_2 предлагается рассчитывать на основе предварительного определения для каждого из слов, входящих в перспективные словосочетания, следующей частной характеристики [Гулин, 2006]:

$$p_{OnTopic} \approx 1 - exp(-1.5 * \frac{CF}{Q}) \quad (7)$$

где $p_{OnTopic}$ – вероятность того, что анализируемое слово не случайно попало в документ, а имеет непосредственное отношение к его тематике; CF – число вхождений слова в коллекцию текстов; Q – количество текстов в корпусе.

Значение коэффициента K_2 можно определить следующим образом:

$$K_2 = max(p_{OnTopic}(w_1), p_{OnTopic}(w_2), \ldots, p_{OnTopic}(w_n)), \quad (8)$$

где $w_1 \ldots w_n$ – слова из словосочетания, претендующего на роль концепта в онтологии; n – количество слов в словосочетании.

Предлагается обобщенный алгоритм нахождения концептов для проектируемой онтологии:

1) Поиск концептов, состоящих из одного слова;
2) Поиск в текстах словосочетаний и расчет по формуле (3) силы связи между словами в словосочетаниях;
3) Исключение из списка тех словосочетаний, сила связи между любой из пар слов в которых меньше характеристики $min\,BF$;
4) Расчет характеристики $Score_1$ для всех текстов по всем однословным претендентам на концепт и по всем словосочетаниям-претендентам по формуле (5);
5) Формирование списков релевантных текстов для каждого словосочетания и слова;
6) Исключение из списка тех словосочетаний и слов, для которых количество релевантных документов меньше, чем $minQ$, и ранжирование по количеству релевантных текстов.

Надо отметить, что средства программной реализации процедуры автоматического построения онтологии по инициирующему запросу эксперта, должны содержать опцию, позволяющую эксперту определять связи между выделяемыми концептами онтологии.

Поиск связей между концептами формируемой онтологии

Рассмотрим задачу нахождения связей между концептами типа «отношение». Выявление таких отношений позволяет при работе с онтологией максимально использовать заложенные в нее знания.

Существуют два различных подхода к поиску связей для будущей онтологии:

1) Поиск связи для двух уже существующих концептов онтологии;
2) Поиск слов, вероятность использования которых в качестве связей онтологии высока, с последующим выбором концептов, которые может соединить в будущей онтологии данная связь.

Рассмотрим возможность реализации второго подхода с помощью алгоритма Гинзбурга [Гинзбург, 1998]. Предлагаемую процедуру поиска связей можно представить следующим образом:

- Поиск претендентов на использование в качестве связей будущей онтологии;
- Формирование по алгоритму Гинзбурга контекстного множества слов-претендентов;
- Перебор пар слов из контекстных множеств для формирования триад «слово из контекстного множества (либо словосочетание, в котором главным является такое слово) + слово-претендент на связь + слово из контекстного множества (либо словосочетание, в котором главным является такое слово)»;
- Закрепление связей в проектируемой онтологии.

Рассмотрим отдельные этапы этой процедуры. Для выделения в текстах претендентов-слов, которые могут быть использованы в качестве связей, выделим в тексте слова W_i с наиболее высоким коэффициентом $K(w_i)$, рассчитываемым по следующей формуле:

$$K(w_i) = \frac{1}{|R_1 - R_2|} * \frac{Q(w_i)}{N_{max} - N_{min}}, \quad (9)$$

где R_1, R_2 – позиции в упорядоченном списке наиболее важных слов, отобранных по коэффициенту $K(w_i)$ и по коэффициенту TF/IDF соответственно; $Q(w_i)$ – количество текстов, в которых присутствует слово W_i; N_{max}, N_{min} – наибольшее и наименьшее количества вхождений слова W_i в текст из корпуса соответственно.

В соответствии с (9), для слова, претендующего на использование в качестве связи между концептами, важно, чтобы оно встречалось в большинстве текстов корпуса, и при этом было максимально равномерно распределено в каждом тексте корпуса. Для каждого из найденных слов, которые будут связывать концепты проектируемой онтологии, составляется контекстное множество. Очевидно, что если в онтологии необходимо связать два понятия, то оба эти понятия должны входить в контекстное множество, составленное для слова-связи. Соответственно концепты, которые будет связывать найденное слово-претендент, нужно искать в контекстном множестве этого слова. Также стоит отметить, что поскольку концепты онтологии могут состоять не только из одного слова, а и быть словосочетаниями, то необходимо разработать метод определения концептов-словосочетаний, которые будет объединять рассматриваемая связь. Рассмотрим варианты «связок» для проектируемой онтологии. В частности, связками могут быть: слово, которое обозначает связь между двумя концептами, или же два концепта, которые могут представлять собой слово либо словосочетание. Связка может иметь вид «слово№1 – связь – слово№2», при этом, если концепт представлен словосочетанием, то в связку вносится главное слово. Для связок «слово – связь – слово» в соответствии с рассматриваемым подходом необходимо осуществить ранжирование возможных претендентов на использование в качестве связных концептов. Очевидно, что вероятность попасть в онтологию выше у тех связок, сила связи между концептами которых больше. Введем частную характеристику $min\,BF_1$:

$$min\,BF_1 = K_3 * max(BF)_1 \quad (10)$$

где K_3 – пороговый коэффициент, позволяющий редуцировать множество возможных связок (рекомендуемый диапазон значений K_3 : 0.3-0.5); $max(BF)_1$ – общее количество связок между концептами-претендентами.

К связкам между концептами будущей онтологии могут быть отнесены лишь те связки, сила связи между элементами которой (в смысле (3)) не меньше, чем $min\,BF_1$.

Если концепт или концепты в связке представлены в виде словосочетаний, то выбрав из словосочетания главное слово, будем рассматривать связку, как связку с концептами из одного слова.

Задача выделения главного слова в словосочетании не является тривиальной и ее однозначное решение не всегда может быть получено даже с помощью стандартного морфологического анализатора. В связи с этим предлагается следующий подход к упрощенному анализу словосочетаний:

- На этапе фильтрации претендентов-концептов выделяем только именные словосочетания;
- С помощью модуля синтаксического анализа находим в словосочетании имя существительное (если их в словосочетании несколько, то выбираем слово, которое стоит в именительном падеже), которое будем считать главным словом в словосочетании.

Для каждой связки определим коэффициент CB_force:

$$CB\_force = B\_force(w_1, v_1) * B\_force(w_1, link) * \ldots * B\_force(w_n, link) * \\ * B\_force(v_1, link) * \ldots * B\_force(v_m, link) \quad (11)$$

где w_1 – главное слово из первого словосочетания; v_1 – главное слово из второго словосочетания; $w_2 \ldots w_n$ – слова из первого словосочетания; $v_2 \ldots v_m$ – слова из второго словосочетания; $link$ – слово, претендующее на роль связи в формируемой онтологии.

Очевидно, что в онтологии должны быть закреплены лишь связки с высокими значениями коэффициента CB_force. Таким образом, в предлагаемой процедуре учитывается не только смысловая связь между главными словами словосочетаний, но и то, насколько словосочетание в целом связано с каждым словом, претендующим на роль связи между концептами.

Программная реализация

Для реализации предложенного подхода был разработан программный модуль «Concept-Ont», который может эффективно использоваться для задач семантического анализа электронных текстов и формирования множества связных концептов при автоматическом синтезе онтологий.

Модуль поддерживает следующие функции:

1) Поиск в корпусе текстов однословных концептов (рис. 1);
2) Поиск в корпусе текстов многословных концептов (рис. 2);
3) Поиск связок для будущей онтологии (рис. 3).

Тестовое моделирование «Concept-Ont» проводилось для корпуса текстов по узкой тематике из электронной библиотеки аннотаций авторефератов диссертационных работ и методических указаний Харьковского национального университета радиоэлектроники.

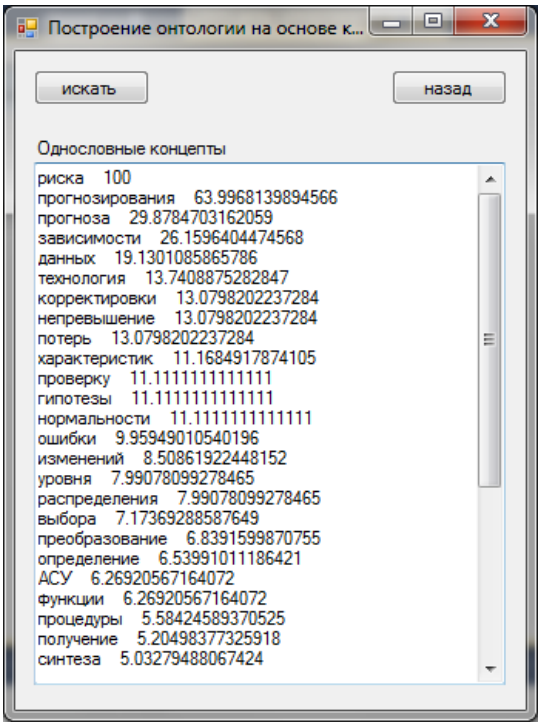


Рисунок 1. Окно поиска однословных концептов

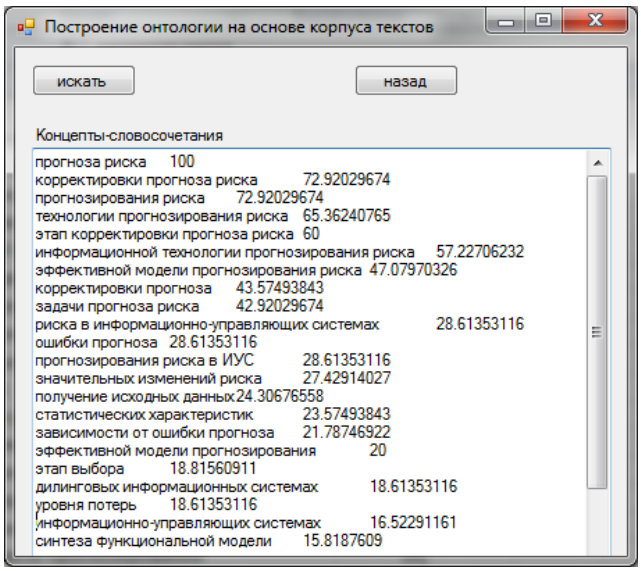


Рисунок 2. Окно поиска концептов-словосочетаний

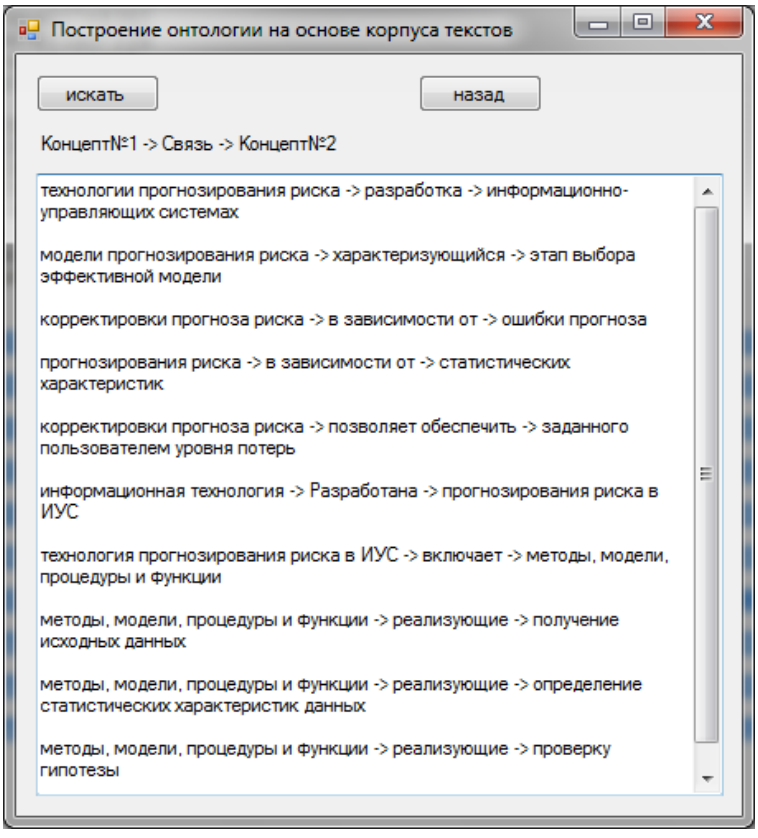


Рисунок 3. Окно поиска связок «концепт№1 + связь + концепт№2»

Для оценки эффективности разработанных методов автоматического формирования связных концептов онтологической модели было проведено сравнение полученных результатов с результатами создания онтологии экспертом предметной области на основе представленных текстов. Следует отметить высокую степень совпадения множеств связных концептов онтологических моделей, полученных как экспертом, так и при автоматическом формировании модулем «Concept-Ont» (в частности, для предметной области «Методы вычислительного интеллекта» такое совпадение составило 97%).

Заключение

Проведенные исследования позволяют сделать вывод, что важным этапом автоматического построения онтологий является формирование связных концептов. Модификация и программная реализация методов нахождения концептов с учетом связей между ними позволили повысить возможности автоматического создания онтологий. При этом особое внимание следует уделить задачам ранжирования однословных/многословных концептов и выявления связей типа «отношения» между ними. При проведении дальнейших исследований целесообразно усовершенствовать предложенный метод, дополнив его анализом более сложных типов онтологических связок.

Литература

[Jinxi Xu, 1996] Jinxi Xu and W. Bruce Croft. Query expansion using local and global document analysis. In Proc. of the SIGIR'96, pp. 4-11, 1996.

[Артемьева, 2006] Мультидисциплинарная система управления информационными ресурсами различных уровней общности. И. Л. Артемьева [и др.]. Проблемы управления : научно-техн. журн. – 2006. - № 4. – С. 64-68.

[Артемьева, 2008] Артемьева И.Л. Онтологии предметных областей и их использование при создании программных систем. Труды симпозиума "Онтологическое моделирование". Звенигород, 19-20 мая 2008. – М.: ИПИ РАН, 2008. – С. 83-113.

[Воронина, 2010] Воронина И.Е. Алгоритмы определения семантической близости ключевых слов по их окружению в тексте. И.Е. Воронина, А.А. Кретов, И.В. Попова. Вестник ВГУ – 2010 – №1 – с. 148-149.

[Гинзбург, 1998] Гинзбург Е. Л. Идиоглоссы: проблемы выявления и изучения контекста. Е. Л. Гинзбург. Семантика языковых единиц: Доклады VI Международной конференции. Т. I., М., 1998. – С. 26–28.

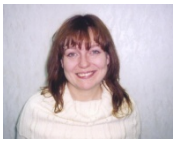
[Гулин, 2006] Андрей Гулин, Михаил Маслов, Илья Сегало Алгоритм текстового ранжирования Яндекса на РОМИП-2006 // http:// romip.ru/ romip2006/03/ 2006.

[Зябрев, 2010] Зябрев И.Н., Пожарков О.В., Пожаркова И.Н. Использование спектральных характеристик лексем для улучшения поисковых алгоритмов. Труды РОМИП 2010. Казань: Казанский ун-т. С. 40-48.

[Чалая, 2011] Чалая, Л. Э. Меры важности концептов в семантической сети онтологической базы знаний. Л.Э. Чалая, Ю. Ю. Шевякова, А. Ю. Шафроненко. Матеріали другої міжнар. наук.-техн. конф. «Сучасні напрями розвитку інформаційно-комунікаційних технологій та засобів управління». – Киев : КДАВТ, 2011. – С. 51.

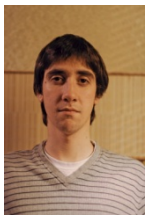
[Чалая, 2012] Чалая, Л. Э. Метод автоматического построения онтологии с древовидной структурой концептов. Л.Э. Чалая, А.В. Чижевский, Ю.Ю. Шевякова. Материалы Международной научно-технической конференции «Информационные системы и технологии» ИСТ-2012. – Морское: 2012. – С.75.

Информация об авторах



Лариса Чалая *– к.т.н., доцент, доцент Харьковского национального университета радиоэлектроники; пр. Ленина 14, 61166, Харьков, Украина; e-mail: kovalivnich@yahoo.com*

Основные области научных исследований: искусственный интеллект, информационный поиск, обработка естественно-языковой информации, фракталы.



Антон Чижевский *– аспирант Харьковского национального университета радиоэлектроники; пр. Ленина 14, 61166, Харьков, Украина; e-mail: chij7@mail.ru*

Основные области научных исследований: искусственный интеллект, обработка естественно-языковой информации, онтологический инжиниринг

CREATING SET OF RELATED CONCEPTS FOR AUTOMATIC ONTOLOGY SYNTHESIS

Larysa Chala, Anton Chizhevskyi

Abstract: *Proposed in this paper the method for finding related concepts in domain can improve the quality of automatically generated ontologies. The method can be efficiently used for tasks of semantic search systems analysis of electronic texts and automatic creation of ontological domain models.*

Keywords: *related concepts, ontology, morphological properties of words, automatically generated ontologies, domain*