УДК 004.82

д.т.н., доц. **Хайрова Н.Ф.** (НТУ «ХАИ»)
Петрасова С.В. (НТУ «ХАИ»)
д.т.н., проф. **Ленков С.В.** (ВИКНУ)

МЕТОД АВТОМАТИЧЕСКОЙ ИДЕНТИФИКАЦИИ СЕМАНТИЧЕСКИХ КОРРЕЛЯЦИЙ ТЕРМИНОВ ГЛОССАРИЯ

В работе предлагается метод автоматической идентификации концептов и их отношений для построения семантической сети предметной области. Рассматриваются семантические корреляции терминов глоссария с точки зрения возможности экстракции и идентификации концептов и их отношений. Предложенная математическая модель позволяет выделить классы толерантности терминов за счет факторизации пространства концептов. Для формализации категорий межконцептуальных отношений узлов семантической сети предлагается использовать диапазон значений коэффициента семантической близости и шаблоны лексических последовательностей.

Ключевые слова: идентификация отношений концептов, семантическая сеть, глоссарий, классы толерантности, межконцептуальные отношения, семантическая близость.

Введение. Все большее влияние на эффективность работы современных интеллектуальных информационных систем оказывают полнота и точность баз знаний, на которых основываются данные системы. При этом значение имеют не только и не столько выбранные модели представления знаний, но и в большей степени наличие возможности их автоматического динамического наполнения, изменения и настраивания на добавляющиеся и изменяющиеся предметные области. С этой точки зрения одной из наиболее перспективных моделей представления знаний остаются различные реализации семантических сетей, архитектура которых предусматривает точную формализацию смысла концептов через отношения между ними. Семантические сети (СС), представляющие универсальную структуру выражения и аккумуляции (наращивания) семантики, предназначены для явной и наглядной систематизации знаний о предметной области, в виде сетевой структуры, основанной на теории графов.

В качестве основного источника знаний для автоматического динамического наполнения и расширения СС могут быть использованы тексты естественного языка. Использование текстов в качестве информационной базы СС требует выделения в них классов элементов, играющих определенно уверенную функциональную роль в представлении знаний. Такими классами являются класс терминологических понятий и класс отношений, которые определяют соответственно множество узлов и множество дуг семантической сети.

Под терминологическим понятием, представляющим собой концепт (concept) предметной области (ПО), понимается совокупность суждений о каком-либо объекте, отражающем его сущность. Результатом познания этого объекта является мысль, в которой обобщаются и выделяются предметы некоторого класса по уверенным, общим и специфическим в совокупности для них признакам. Под отношением (relation) понимается философская категория или научный термин, обозначающий любое понятие, реальным коррелятом которого является определенная связь двух и более концептов [1].

Такое определение терминологических понятий и отношений предметной области позволяет рассматривать не синтаксические связи языковых единиц в тексте, а устойчивые семантические связи между терминологическими понятиями, имеющими энциклопедический характер.

Наиболее емко «энциклопедические знания» отображены в глоссариях, где словарные статьи представляют собой тексты с концентрированной смысловой нагрузкой. В информационном ресурсе глоссария явным образом выделяются концепты, соответствующие понятийному аппарату ПО, и не явно, но насыщено и концентрировано представляются семантические отношения между понятиями ПО. Семантическая сеть,

построенная на основе структурированных текстов глоссария, позволит не только наиболее полно и точно выделять концепты и формализовать семантические отношения между концептами ПО, но и автоматически дополнять СС, в случае изменения и дополнения глоссариев.

Постановка задачи исследования. Основной проблемой автоматической экстракции из глоссариев концептов и семантических отношений между концептами, которые неявно представлены в естественно-языковых конструкциях, является сложность формализации категориального аппарата семантической сети, которая использует данные элементы. Концепты терминов, представляющих узлы семантической сети, обладают некоторыми семантическими корреляциями, выражающими определенную общность содержания или сходства обозначаемых явлений [2]. Связи между значениями слов, связанных в семантической сети дугами, различаются по степени общности.

В данном исследовании для решения задачи формального выделения категорий межконцептуальных отношений предлагается использовать метод компонентного анализа. Данный метод основывается на предположении о том, что обладающие семантическими корреляциями концепты имеют определенную общность содержания, выражающую некоторое сходство обозначаемых явлений или понятий [2, 3]. При этом виды связей между значениями слов, связывающих концепты семантической сети, различаются по степени общности или эквивалентности. Используемый метод идентификации отношений толерантности и эквивалентности позволяет разделить корреляции концептов СС по степени семантической близости, классифицируя отношения в соответствии с традиционными типами: гиперонимия, гипонимия, холонимия, меронимия и семантическая эквивалентность.

Математическая модель идентификации отношений толерантности и эквивалентности терминов глоссария. Для построения логической схемы выделения корреляционно зависимых концептов ПО вводится метрическое пространство лингвистических смысловых единиц Θ, определяемое как множество терминов глоссария T, на котором грамматические правила задают отношения между единицами, выступающими ограничениями для корректных синтаксических структур [4].

Для определения метрики пространства используем меру семантической близости *f(t', t'')* между двумя лингвистическими единицами *t' и t''*, выражаемую через отношение теоретико-множественного пересечения и объединения множеств терминов дефиниций:

$$f(t',t'') = \frac{2\,|\,d_1 \bigcap d_2\,|}{|\,d_1\,| + |\,d_2\,|}, \qquad (1)$$

где $d_1 \bigcap d_2$ — общие термины дефиниций глоссария, а $|\,d_1\,| + |\,d_2\,|$ — все термины дефиниций d_1 и d_2; под термином в данном контексте мы понимаем концепт из глоссария, взятый в его канонической форме.

Дефиницию, включающую термины глосария $t \in \Theta$, обозначим как d. Пара элементов $(t, d) \in (\Theta, \Omega)$ представляет собой один термин и один фрагмент связного текста дефиниции глоссария (фраза, предложение, полное определение), где Θ — пространство лингвистических единиц рассматриваемого глоссария T, а Ω — пространство рассматриваемых фрагментов текстов глоссария.

Будем говорить, что две лингвистические единицы связаны в одном семантическом поле и писать $(t_i, d_i) \sim (t_j, d_i)$, если только $F(t_i, d_i) = F(t_j, d_i)$. Например, *F("компилятор", d_1[1]); F("транслятор", d_1); F ("компилятор", d_1) = F("транслятор", d_1)*.

Можно показать, что отношение ~, устанавливаемое между терминами *t* и элементами связного текста *d*, выражает толерантность и факторизует пространства лингвистических смысловых единиц Θ и исследуемых связных текстов Ω, разбивая их на классы толерантности.

---

[1]компьютерная программа, преобразующая текст программы пользователя, написанный на языке программирования высокого уровня в исходный для перевода в машинный код процессора соответствующей платформы;

Толерантностью называется отношение, обладающее свойствами рефлексивности и симметричности [5]. Можно показать, что отношение $(t_i, d_i) \sim (t_j, d_i)$ является рефлексивным отношением, т. е. один термин глоссария в одном своем сигнификативном значении связан сам с собой, и симметричным, один термин глоссария в одном своем сигнификативном значении связан с другим (в одном из его значений) и одновременно второй термин связан отношением ~ с первым (в вышеназванных значениях):

$$(\text{“транслятор”}, d_2^2) \sim (\text{“интерпретатор”}, d_2) \leftrightarrow F(\text{“транслятор”}, d_2) = F(\text{“интерпретатор”}, d_2) \equiv F(\text{“интерпретатор”}, d_2) = F(\text{“транслятор”}, d_2) \leftrightarrow \leftrightarrow (\text{“интерпретатор”}, d_2) \sim (\text{“транслятор”}, d_2).$$

Пространство толерантности S_p, где p – число классов толерантности, состоящее из множеств номеров вида $N=\{n_1, n_2, ..., n_k\}$, при этом все $n_i \leq p$, причем элементы (t_i, d_i) и (t_j, d_i) толерантны, если они содержат общий номер. Множество K_h является классом толерантности, если K_h состоит из всех множеств вида $\{i, n_1, ..., n_k\}$ и число элементов множества K_h равно 2^{p-1} – число всех подмножеств множества из оставшихся $p-1$ номеров.

Смысл этого утверждения состоит в том, что отношение $(t_i, d_i) \sim (t_j, d_i)$ выполняется тогда и только тогда, когда существует класс K_i, содержащий одновременно (t_i, d_i) и (t_j, d_i). Если $(t_i, d_i) \sim (t_j, d_i)$, то (t_i, d_i) и (t_j, d_i) содержат некоторый общий номер h, и тем самым входят в класс K_h [5].

Частным случаем отношения толерантности является отношение эквивалентности. Чтобы показать, что отношение ~, устанавливаемое между терминами глоссария t и элементами связного текста d, выражает эквивалентность и факторизует пространства лингвистических смысловых единиц Θ и исследуемых связных текстов Ω, разбивая их на классы эквивалентности, достаточно показать, что отношение ~ является не только рефлексивным и симметричным, но и транзитивным [6].

Например, отношение ~ является транзитивным отношением, если один термин глоссария в одном из своих значений имеет тот же сигнификативный смысл, что и второй термин в одном из своих значений, и второй термин в уже обозначенном значении имеет тот же сигнификативный смысл, что и третий в одном из своих смыслов, и тогда первый термин в определенном сигнификативном значении связан с третьем:

$(\text{“компилятор”}, d_1) \sim (\text{“транслятор”}, d_1)$ и $(\text{“транслятор”}, d_1) \sim (\text{“интерпретатор”}, d_1) \leftrightarrow F(\text{“компилятор”}, d_1) = F(\text{“транслятор”}, d_1) = F(\text{“интерпретатор”}, d_1)$.

Данное отношение эквивалентности позволяет организовать различные пары терминов и фрагментов связных текстов, включающих данные единицы, (t, d), в классы эквивалентности, которые определяют один и тот же сигнификативный смысл, тем самым, позволяя факторизовать пространство концептов, выражаемых знаками лингвистических смысловых единиц, на классы синонимичных в каком-то из своих смыслов концептов [7].

Эвристическая оценка семантической корреляции концептов. Для определения меры семантической близости (1), при которой семантическая корреляция терминов формализуется отношениями эквивалентности и толерантности, исследовалась выборка из тысячи терминов глоссария по информационным технологиям [8]. В результате эксперимента было автоматически выявлено 5000 семантически близких концептов (табл. 1) и определен коэффициент семантической близости между ними.

[2]компьютерная программа, выполняющая преобразование программы, представленной на одном языке программирования, в программу на другом языке (перевод программы во внутренний язык ее процессора)

Таблица 1

Семантически близкие концепты

Коэффициент семантической близости	Количество связанных концептов (%)	Пример семантически близких концептов
0,81-0,9	5 (0,1%)	конкурентное программирования ~ параллельное программирование
0,51-0,8	95 (1,9%)	истинная погрешность ~ абсолютная погрешность
0,36-0,5	300 (6%)	математическое обеспечение ~ программное обеспечение
0,28-0,35	750 (15%)	размерность ~ измерение
0,21-0,27	2250 (45%)	реентерабельна программа ~ повторное использование
0,05-0,2	1600 (32%)	среда ~ мониторинг

Используя математическую модель идентификации отношений эквивалентности терминов глоссария можно осуществить разбиение терминов только на классы эквивалентности. Для того чтобы получить так же разбиение терминов на классы толерантности, которые объединяют элементы по некоторому общему интегральному семантическому признаку и различаются, по крайней мере, по одному дифференциальному признаку, используется эвристический анализ коэффициентов семантической близости. Такой анализ показывает, что концепты терминов, связанные отношением толерантности, попадают в один класс, при значении коэффициента семантической близости, вычисляемом по формуле (1), 0,28-0,35. Именно к этой категории семантических отношений относятся отношения гиперонимии, гипонимии, холонимии, меронимии. Тогда как семантическими эквивалентами могут считаться те концепты, мера семантической близости которых входит в диапазон от 0,36 до 1. Семантической связностью терминов, у которых мера семантической близости ниже 0,28, можно при построении СС пренебречь.

Например, рассмотрим отношение толерантности на примере концептов t_1=*"диалоговое меню"* и t_2=*"графический интерфейс пользователя"*. Определенная мера семантической близости *f*:

$$f(t_1, t_2) = \frac{2 \times 3}{10 + 11} = 0{,}29$$

Полагая, что концепты t_1 и t_2 связаны отношением толерантности, покажем, что данное отношение $(t_1, d_1) \sim (t_2, d_2)$ является рефлексивным и симметричным:

(“диалоговое меню”, d_1) ~ (“графический интерфейс пользователя”, d_2) ↔
↔ F (“диалоговое меню”, d_1) = F (“графический интерфейс пользователя”, d_2) ≡
≡ F (“графический интерфейс пользователя”, d_2) = F (“диалоговое меню”, d_1) ↔
↔ (“графический интерфейс пользователя”, d_2) ~ (“диалоговое меню”, d_1).

Определим коэффициент семантической близости (1) терминов: t_1=*"программист"*, t_2=*"разработчик"* и t_3=*"девелопер"*, используя дефиниции (t_1, d_3^3); (t_2, d_4^4) и (t_3, d_5^5):

[3] человек, разрабатывающий компьютерные программы;

[4] человек, выполняющий работы по разработке программного обеспечения и информационных систем на различных этапах их жизненного цикла;

[5] человек, вовлеченный во все аспекты процесса разработки программного обеспечения, который в настоящее время рассматривается более широко, чем проектирование программ;

$$f(t_1,t_2)=\frac{2\times 3}{4+10}=0{,}43$$
$$f(t_2,t_3)=\frac{2\times 4}{10+12}=0{,}37$$
$$f(t_1,t_3)=\frac{2\times 3}{4+12}=0{,}38$$

Такие значения коэффициентов семантической близости показывают, что концепты t_1="программист", t_2="разработчик" и t_3="девелопер" связаны отношением эквивалентности. В свою очередь можно показать, что отношение f является транзитивным отношением:

(“программист”, d_3) ~ (“разработчик”, d_3) и *(“разработчик”, d_3) ~ (“девелопер”, d_3)* ↔
↔ *F(“программист”, d_3)= F(“разработчик”, d_3)= F(“девелопер”, d_3)*.

Определим меру семантической близости для концептов *(t_1=“интерпретатор”, d_3); (t_2=“компоновщик”, d_6[6])*:

$$f(t_1,t_2)=\frac{2\times 1}{10+14}=0{,}09$$

Значение коэффициента $f(t_1,t_2)$ ниже диапазона 0,28-1, из чего следует, что рассматриваемые концепты “*интерпретатор*” и “*компоновщик*” не являются семантически близкими концептами.

Метод формализация категорий межконцептуальных отношений СС. Как уже было сказано ранее, отправной точкой создания СС предметной области является выделение категорий межконцептуальных отношений, представленных в СС. Для этого в образовавшемся пространстве терминов, факторизованных в классы толерантности, идентифицируем явные категории семантической корреляции с помощью шаблонов лексических последовательностей, представляемых регулярными выражениями.

Для идентификации отношений принадлежности к единому классу, гиперонимии, гипонимии, холонимии и меронимии узлов семантической сети, на базе факторизованных терминов глоссария, использовались шаблоны лексических последовательностей следующего вида:

$$\mathrm{NN}_1 \rightarrow \mathrm{Rel}_z \rightarrow \mathrm{NN}_2,$$

где NN_1 и NN_2 – связанные концепты, представленные ключевыми словами и словосочетаниями глоссария, Rel_z – лексические цепочки, выражающие отношения z.

Таблица 2

Лексические цепочки, выражающие категории межконцептуальных отношений

z	Тип отношения	Лексические цепочки, Rel_z
class	принадлежность к классу	*"считается", "относится к", "это (является)"*
hypernyms	гиперонимия	*"совокупность", "комплекс", "набор", "семейство"*
hyponyms	гипонимия	*"например", "тип", "вид", "экземпляр"*
HAS-A	холонимия	*"включает", "делится на"*
A-Part-Of	меронимия	*"часть", "элемент"*

Например, концепты t_1="диалоговое меню" и t_2="графический интерфейс пользователя" связаны отношением толерантности. Используя регулярные выражения, идентифицируем в дефинициях d_i и d_{i+1} явные лексические цепочки Rel_z. Такая цепочка представлена лексемой *"элемент"* в дефиниции d_i концепта t_1:

"диалоговое меню" → ”элемент”$_{A\text{-}Part\text{-}Of}$ → "графический интерфейс пользователя"

[6] системная программа, из объектных модулей, сгенерированных компилятором, а также библиотек транслятора - строит выполняемую программу

Программная реализация модели. Семантическую сеть принято рассматривать как структуру, образующуюся в результате композиции отдельных концептуальных графов, отвечающих концептуальным единицам (описанию отдельного понятия). Программная реализация предложенного метода позволяет автоматически выделять концептуальные единицы ПО, представляющие термины, анализируемых глоссариев, и явным образом определять корреляционные семантические отношения, представляющие концептуальные графы данных терминов.

На рис. 1 отображен интерфейс разработанной программы. Пользователь вводит в поле *Концепт* запрос, представляющий тот или иной термин предметной области информационных технологий.

После автоматической обработки подключенного глоссария в поле *Межконцептуальные отношения* выводятся концепты, связанные с запросом пользователя отношениями гиперонимии, гипонимии, холонимии, меронимии и/или отношением принадлежности к единому классу. В таблицах *Эквивалентные концепты* и/или *Толерантные концепты* данного диалогового окна (см. рис. 1) выводятся результаты кластеризации концептов СС за счет факторизации пространства терминов, а именно концепты, связанные отношением эквивалентности и/или толерантности, с указанием коэффициента семантической близости в порядке возрастания.

Для каждого определенного концепта (отмеченного значком ) возможно отображение его семантических отношений с другими концептами глоссария. Так, например, концепт *"программа"*, связанный с запрашиваемым термином *"компилятор"* отношением принадлежности к классу, в свою очередь является гиперонимом для таких концептов глоссария, как *"компилятор"*, *"интерпретатор"*, *"транслятор"*, *"компоновщик"*.

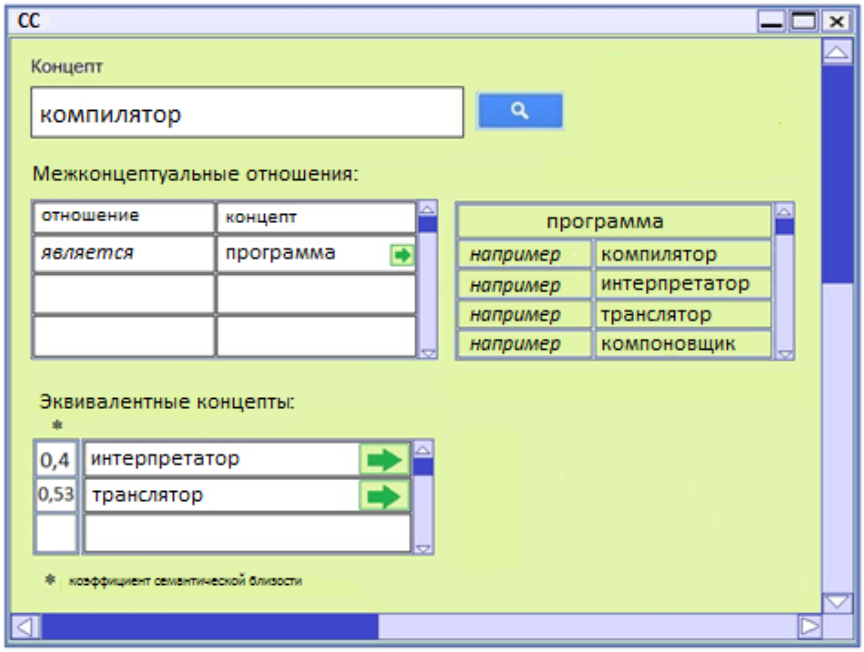

Рис. 1. Программная реализация

Выводы. В работе рассмотрен метод автоматического построения СС, основанный на использовании глоссария как естественно-языкового текста, наиболее полно концентрирующего знания ПО. В качестве узлов создаваемой СС экстрагируются термины глоссария, а для выделения семантических отношения используется математическая модель идентификации отношений толерантности и эквивалентности терминов глоссария. Модель позволяет кластеризовать концепты СС за счет факторизации пространства терминов, используя в качестве основания классификации смысловую близость терминов. Категориальным значением синонимичных лингвистических единиц при этом выступает единое смысловое поле рассматриваемых дефиниций. Проведенная эвристическая оценка значений коэффициента семантической близости позволяет определить диапазон

допустимых значений коэффициента в отношениях семантической толерантности и семантической эквивалентности терминов глоссария.

Узлы создаваемой СС представлены терминами глоссария, а размеченные дуги определяются смысловыми отношениями — принадлежность к классу, гиперонимия, гипонимия, холонимия и меронимия. Явное выделение данных категорий семантических отношений осуществляется с помощью регулярных выражений, представленных шаблонами лексических последовательностей.

Программная реализация модели показала при экспериментальном построении СС на базе глоссария по информационным технологиям объемом около тысячи терминов приемлемость использования данного подхода.

ЛИТЕРАТУРА:

1. Философия: энциклопедический словарь / ред. А. А. Ивина. – М. : Гардарики, 2004. – 1072 с.

2. Кобозева И. М. Лингвистическая семантика: Учебное пособие. / И. М. Кобозева. – М. : Эдиториал УРСС, 2000. – 352 с.

3. Широков В. А. Інформаційна теорія лексикографічних систем: моногр. / В. А. Широков. – К. : Довіра, 1998. — 331 с.

4. Хайрова Н. Ф. Концептуальная схема идентификации смысла лингвистических единиц / Н. Ф. Хайрова // Збірник наукових праць Військового інституту Київського національного університету імені Тараса Шевченка. – К. : ВІКНУ, 2013. – Вип. № 39. – С.217-223

5. Шрейдер Ю. А. Равенство, сходство, порядок / Ю. А. Шрейдер. – М. : «Наука», 1971. – 256 с.

6. Бондаренко М. Ф. Теория интеллекта: учебник / М. Ф. Бондаренко, Ю. П. Шабанов-Кушнаренко. – Харьков : Комп. СМИТ, 2007. – 576 с.

7. Хайрова Н. Ф. Определение семантической близости на основе когнитивного подхода. / Н. Ф. Хайрова, Н. В. Шаронова, Н. В. Борисова // Бионика интеллекта: науч.-техн. журнал, 2013.

8. Толковый словарь по информатике / Г. Г. Пивняк, Б. С. Бусыгин, М. М. Дивизинюк и др. – Д. : Нац. горн. ун-т, 2010. – 600 с.

Рецензент: д.т.н., проф. Шаронова Н.В., національний технічний університет «ХПІ»

Хайрова Н.Ф., Петрасова С.В., Лєнков С.В.

МЕТОД АВТОМАТИЧНОЇ ІДЕНТИФІКАЦІЇ СЕМАНТИЧНИХ КОРЕЛЯЦІЙ ТЕРМІНІВ ГЛОСАРІЮ

У роботі пропонується метод автоматичної ідентифікації концептів та їх відношень для побудови семантичної мережі предметної області. Розглядаються семантичні кореляції термінів глосарія з точки зору можливості екстракції та ідентифікації концептів та їх відношень. Запропонована математична модель дозволяє виявити класи толерантності термінів за рахунок факторизації простору концептів. Для формалізації категорій міжконцептуальних відношень вузлів семантичної мережі пропонується використовувати діапазон значень коефіцієнта семантичної близькості та шаблони лексичних послідовностей.

Ключові слова: ідентифікація відношень концептів, семантична мережа, глосарій, класи толерантності, міжконцептуальні відношення, семантична близькість.

Prof. Khairova N.F., Petrasova S.V., Prof. Lenkov S.V.

THE METHOD OF THE AUTOMATIC IDENTIFICATION OF SEMANTIC CORRELATIONS BETWEEN GLOSSARY TERMS

The paper proposes the method of the automatic identification of concepts and their relations for building a domain semantic network. We consider semantic correlations between glossary terms from the viewpoint of possibility of the extraction and identification of concepts and their relations. The proposed mathematical model allows to identify classes of tolerance of the terms due to the factorization of the space of concepts. To formalize the categories of relations between semantic network nodes a range of values of the coefficient of semantic proximity and patterns of lexical sequences are encouraged to use.

Keywords: identification of relations of concepts, semantic network, glossary, tolerance classes, relations of concepts, semantic proximity.