Valery Mihaylovich Nemchinov, *PhD. tech. Sciences, ProfessorNational research nuclear University"Moscow engineering physics Institute"*

This article describes a method of the development measuring devices using the unified hardware and software platform. This method allows for a fast and simple development of measuring instruments.

Keywords: Measuring means, sensor, programming, embedded system, microcontroller, unification.

УДК 004

МОДЕЛЬНОЕ ПРЕДСТАВЛЕНИЕ МЕХАНИЗМОВ ФОРМИРОВАНИЯ ПРОБЛЕМНО-ОРИЕНТИРОВАННЫХ СЕМАНТИЧЕСКИХ ПОЛЕЙ

Александр Сергеевич Сигов, *академик РАН, президент,*
E-mail: sigov@mirea.ru,
Валентин Викторович Нечаев, *академик РАЕН, зав. каф.,*
E-mail: nechaev@mirea.ru,
Всеволод Михайлович Трофименко, *аспирант,*
E-mail: trofsev@mail.ru,
Московский государственный технический университет радиотехники, электроники и автоматики,
https://www.mirea.ru

Рассматривается один из возможных подходов к задаче построения семантического поля в некоторой проблемно-ориентированной области знаний путем выделения лексических единиц из слабоструктурированных или неструктурированных текстовых документов. В основу построения модели семантического поля заложены некоторые научные факты из когнитивной психологии и психолингвистики.

Ключевые слова: семантическое поле, бионический подход, интеллектуальный анализ данных, эвристические методы, эвристическое моделирование.

Исследование выполнено федеральным государственным бюджетным образовательным учреждением высшего профессионального образования «Московский государственный технический университет радиотехники, электроники и автоматики» (МГТУ МИРЭА) за счет гранта Российского научного фонда (проект № 14-11-00854)

Введение

Интенсивный рост объемов различного вида (неоднородных) информационных ресурсов, многообразные потребности общества и личности в оперативной семантической обработке таких ресурсов и, как следствие, проблема больших данных (Big data), настоятельно требуют создания новых информационно-аналитических систем, ориентированных на автоматизацию решения задач, соответствующих возникающим потребностям. Одним из перспективных подходов к решению по-

А.С. Сигов

В.В. Нечаев

добного рода проблем является бионический.

Моделирование функций головного мозга человека (ГМЧ) как информационно-аналитической системы и техническая реализация подобной модели, с точки зрения авторов, дает возможность разработать методы и алгоритмы обработки семантической информации с учетом релевантности и пертинентности по отношению к запросам пользователей. Очевидно, что создание адекватной бионической модели информационно-аналитической функции ГМЧ (ИАФ ГМЧ) чрезвычайно сложная задача, включающая модели восприятия, осознания, осмысления и означивания информации на уровне понятий и их ансамблей. Создание таких моделей и их объединение в единую систему дает возможность порождать (генерировать) семантические последовательности, ассоциативные цепочки, а также соответствующие многомерные поля. Модельная реализация моделирующей функции ГМЧ дает возможность целенаправленно формировать информационные образы на основе существующих баз знаний. Создание модели прогностической функции ГМЧ на основе соответствующих когнитивных схем может способствовать разработке принципиально новых алгоритмов прогнозирования на основе релевантных и пертинентных знаний. Именно к описанному выше направлению исследований и относится данная статья.

Формирование проблемно-ориентированного семантического поля

На подготовительном этапе для дальнейшего формирования семантической сети понятий предметно-ориентированной области необходимо выделить текстовые документы, соответствующие рассматриваемой предметной области. Данную операцию можно выполнить либо «вручную», благодаря экспертам предметно-ориентированной области, которые отберут необходимые документы, либо в автоматическом (или полуавтоматическом) режиме – на основании ключевых слов текстовых документов. Полуавтоматический режим предполагает перепроверку итогового списка документов экспертом предметно-ориентированной области. При этом ключевые слова могут быть выделены на основании статистических законов Ципфа-Мандельброта и выводов, которые можно сделать из этих законов. Кроме того, автоматизированная система может воспользоваться ключевыми словами, которые были выделены в самих текстовых документах.


В.М. Трофименко

После формирования списка документов должно происходить автоматическое выделение лексических единиц из рассматриваемых текстов. В ходе анализа могут быть выделены не только сами лексические единицы (слова или словосочетания), но и связи между ними.

В данной статье рассматриваются вопросы выделения подобных связей и построения на основании выделенных лексических единиц, а также связей между ними семантического поля некоторой проблемно-ориентированной области.

Выделение связей между понятиями

Для установления связей между элементами некоторой сети используется версия метода интеллектуального анализа данных, называемая анализом связей.

Основные этапы метода анализа связей:

1 Выделение объектов рассматриваемой предметной области.

2 Выделение связей между объектами.

3 Составление матрицы взаимодействия (связей).

4 Построение сети для визуализации данных.

5 Анализ построенной сети

6 Определение показателей сети и ее элементов.

Инструментарий реализации анализа связей со временем претерпевал изменения. Условно можно выделить три этапа развития данного метода [1]:

1 Все шесть этапов выполняются экспертом в проблемно-ориентированной области знаний. Данный метод является чрезвычайно трудоемким при рассмотрении больших объемов данных.

2 Появляется возможность автоматизации построения сети для соответствующих матриц взаимодействия. Однако ввод данных по прежнему необходимо осуществлять «вручную». Процедуры анализа данных также требуют активного участия эксперта проблемно-ориентированной предметной области.

3 Возникает возможность автоматической визуализации связей между объектами. Появляются средства, позволяющие визуально ужимать большие объёмы данных в компактные пучки, что упрощает визуальный анализ данных для сложных моделей. Вычисление показателей сети осуществляется автоматически.

Не смотря на то, что наиболее новые и совершенные методы Анализа связей позволяют вычислять показатели сети и ее элементов в автоматическом режиме, после вычисления всех показателей необходима экспертная оценка для уточнения связей и их корректировки. Ниже приведён детальный анализ этапов метода анализа связей.

Построение матрицы взаимодействия

Удобным и часто используемым методом, используемым при анализе связей, является матрица взаимодействия. Здесь предметом анализа являются возможные взаимосвязи и взаимоотношения между несколькими объектами. Основные этапы построения матрицы взаимодействий [2]:

1 Определение понятия «элемент» и «взаимосвязь» в решаемой задаче.

2 Составление матрицы, в которой каждый элемент может быть сопоставлен с любым другим.

3 На основе объективных данных (исследование, опрос, экспертная оценка) определение, имеется ли взаимосвязь между каждой парой элементов.

В текстовом документе можно выделить такие элементы, как: слова и словосочетания, даты, имена, номера телефонов, адреса и т. д.

В качестве взаимосвязей в текстовом документе могут выступать: ассоциативная связь, связь род-вид, связь причина-следствие, процесс-объект, свойство-носитель свойства, часть-целое, сырье-продукт, административная иерархия, процесс-объект, функциональное сходство, процесс-субъект, антонимия.

Построение матрицы взаимодействия, где в качестве взаимосвязей между словами и словосочетаниями используется ассоциативная связь, рассмотрено ниже.

Ассоциативное связь (ассоциативное отношение) является объединением отношений, не входящих в иерархические отношения или в отношения синонимии. Допускается включать в ассоциативное отношение все виды отношений, кроме синонимии и отношения «род – вид». В матрице «1» обозначает наличие связи, «0» – ее отсутствие.

Таблица 1

Матрица взаимодействия. Ассоциативная связь

№	Наименование	1	2	3	4	5	6
1	Мойдодыр	-	1	1	0	1	0
2	Грязь	1	-	0	0	0	0
3	Начальник	1	0	-	0	0	0
4	Ребёнок	0	1	0	-	0	0
5	Чуковский	1	0	0	1	-	1
6	Айболит	0	0	0	1	1	-

Для того чтобы выделить ассоциативные связи можно воспользоваться методом k-ближайших соседей [3] и методом контрольных списков [4]. Кроме того, исходя из

определения ассоциативной связи, можно сделать вывод о том, что заключение об ассоциативной связи возможно за счёт выделенных ранее связей всех типов, кроме синонимии и отношения «род−вид».

Метод k-ближайших соседей

Человек, сталкиваясь с новой задачей, использует свой жизненный опыт, вспоминает аналогичные ситуации, которые когда-то с ним происходили. О свойствах нового объекта мы судим, полагаясь на похожие знакомые наблюдения. Например, встретив иностранца на улице, мы можем догадаться о его происхождении по речи, жестам и внешности. Для этого необходимо вспомнить наиболее похожего на него человека, происхождение которого известно.

Так, подобно приведенному выше примеру, сходство объектов лежит в основе алгоритма k-ближайших соседей (k-nearest neighbor algorithm, KNN). Алгоритм способен выделить среди всех наблюдений k известных объектов (k-ближайших соседей), похожих на новый неизвестный ранее объект. На основе классов ближайших соседей выносится решение касательно нового объекта. Важной задачей данного алгоритма является подбор коэффициента k – количество записей (количество ближайших соседей), которые будут считаться похожими.

Правило k-ближайших соседей можно сформулировать следующим образом: «Если больше половины свойств одного объекта идентичны свойствам второго объекта, тогда эти объекты можно рассматривать как близлежащие».

Из данного правила можно сделать вывод, что чем больше одинаковых свойств имеют два элемента, тем ближе они расположены друг относительно друга.

Таким образом, с помощью метода k-ближайших соседей можно предложить пользователю некоторый набор возможных ассоциаций (контрольный список), из которых он сможет отсечь ненужные и добавить дополнительные ассоциации. Контрольный список необходимо составлять в зависимости от значения веса понятия (количества повторений понятия в текстовых документах). Сначала необходимо установить связи наиболее важных понятий, потом – всех остальных.

Построение сети для визуализации данных

На основе матрицы взаимодействия, метода k-ближайших соседей и работы пользователя с контрольным списком формируется сеть понятий П, которая в последствии демонстрирует пользователю окружения различных рангов для конкретного выделенного ведущего (доминирующего) понятия, где: П1, П2, П3, П4 – понятия первого ранга, П1.1, П1.2 – понятия второго ранга и т. д. (рисунок 1).

Рассматриваемое ведущее (основное) понятие (рисунок 1) окружают понятия первого ранга (имеющие наибольший коэффициент k). Далее следует окружение второго ранга и т. д. Понятие второго ранга является ассоциацией понятия первого ранга и может быть рассмотрена как вторичная ассоциация применительно к рассматриваемому понятию.

В рамках конкретной предметно-ориентированной области может быть построено множество матриц взаимодействия, которые устанавливают все возможные связи (или большинство связей) рассматриваемой предметной области. На основании данных матриц могут быть построены семантические сети, связывающие между собой понятия предметно-ориентированной области (рисунок 2).

При анализе текстового документа может встретиться понятие, которое ранее не встречалось и не содержится в семантической сети предметно-ориентированной области. Для того, чтобы добавить понятие в семантическую сеть необходимо проверить с помощью метода k-ближайших соседей возможность добавления его в конкретное место сети. Если более половины признаков нового понятия совпадают с признаками включенного в сеть понятия можно рассматривать данные понятия как близлежащие.

При этом возможен случай, когда из текстового документа нельзя установить свойства нового понятия. В этом случае понятие не добавляется в семантическую сеть, а запоминается в отдельной области памяти. В дальнейшем, при проявлении дополнительных свойств понятие может быть сравнено с уже имеющимися понятиями семантической сети и при наличии совпадающих признаков может быть сделан вывод о добавлении нового понятия в сеть.

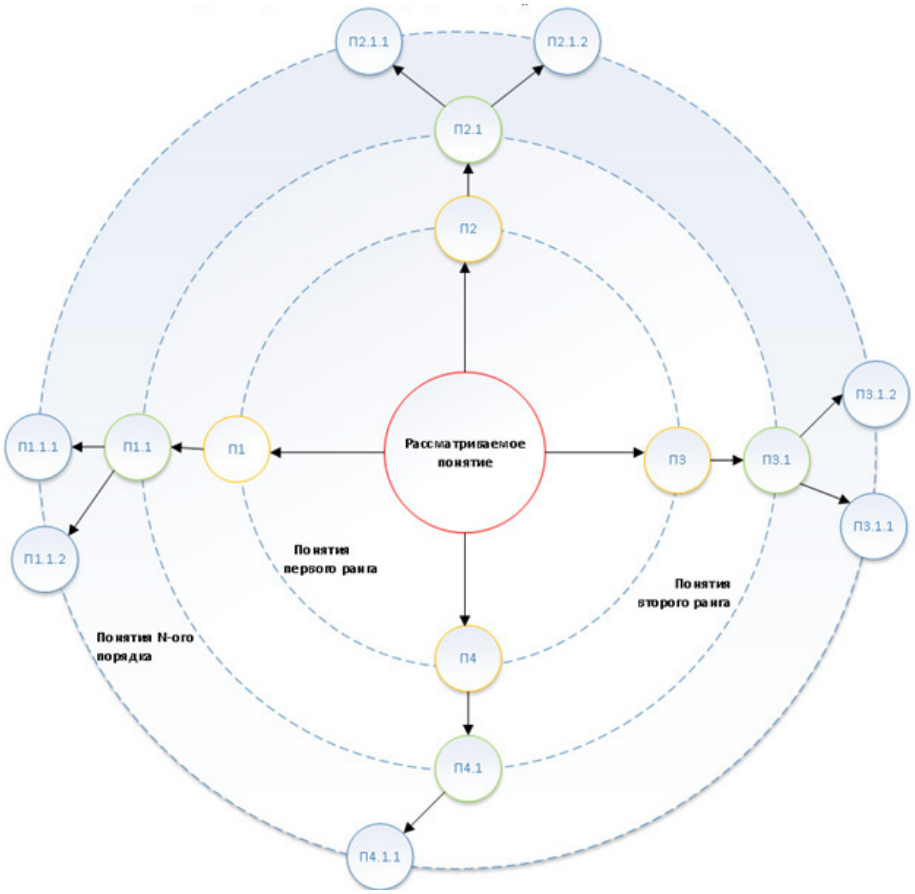

Рисунок 1– Схема семантической сети понятия

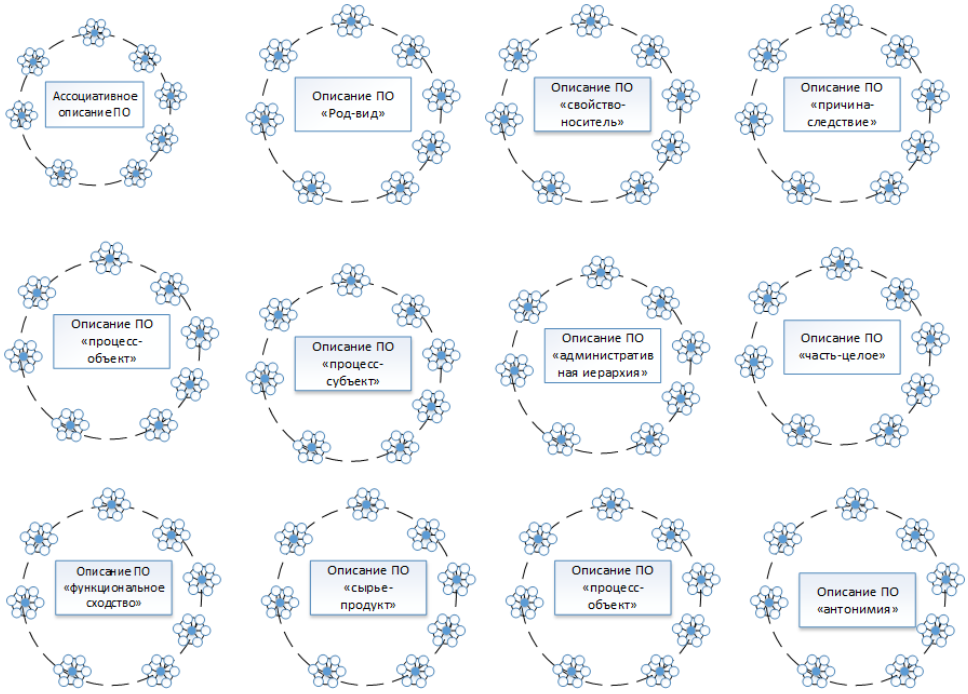

Рисунок 2 – Схема семантической сети предметной области

Обновление семантической сети

Если у двух понятий имеются одинаковые свойства, но их количество не превышает необходимого значения, то их можно добавить в контрольный список для последующего экспертного анализа. Новое понятие может быть перемещено ближе, либо дальше от сравниваемого с ним сетевым понятием, в зависимости от совпадающих свойств соседей сравниваемого понятия. Если понятие встречается не в первый раз, тогда необходимо увеличить его значимость (вес) в рассматриваемой предметно-ориентированной области.

В конечном итоге в центральной части семантической сети должны оказаться наиболее часто встречаемые понятия, затем понятия, которые встречаются меньшее количество раз и т. д.

Вычисление основных показателей сети

Существует ряд общепринятых техник оценки свойств сети и связей между её элементами.

К основным характеристикам сети относятся:

1 ***Размер сети.*** Размер сети может изменяться от минимального значения – 1 (2 вершины в графе) до максимально возможного значения (g-1), где g – количество вершин графа.

2 ***Сетевая плотность*** или сила связанности между объединенными элементами сети. Различают следующие формулы для вычисления сетевой плотности: для неориентированного графа:

$$\Delta = \frac{L}{g(g-1)/2} = \frac{2L}{g(g-1)},$$

для ориентированного графа:

$$\Delta = \frac{L}{g(g-1)},$$

где L – количество наблюдаемых связей в данном графе или подграфе.

3 ***Для определения главного понятия*** в одной сети, например построенной на основе ассоциативных связей, используют формулу степени центральности элемента сети:

$$C_D = d(n_i) = \sum_j x_{ij} = \sum_j x_{ji},$$

где $\sum_j x_{ji}$ – это количество связей между элементами сети [5].

4 ***Для определения главного понятия*** при сравнении нескольких сетей, например ассоциативной и сети "род-вид" используют формулу вычисления нормированной степени центральности элемента сети:

$$C'_D(n_i) = \frac{d(n_i)}{g-1} = \frac{\sum_j x_{ij}}{g-1}.$$

5 ***Для нахождения главной*** (или центральной) сети: используют формулу центральности сети:

$$C_D = \frac{\sum_{i=1}^{g}[C_D(n^*) - C_D(n_i)]}{\max \sum_{i=1}^{g}[C_D(n^*) - C_D(n_i)]}.$$

6 ***Наряду с использованием метода k-ближайших*** соседей, для оценки близости между понятиями используют формулу плотности центральности элемента сети:

$$C_c(n_i) = [\sum_{j=1}^{g} d(n_i, n_j)]^{-1}.$$

Здесь d(ni,nj) – число связей между элементами сети ni и nj. Максимальное значение индекса равно $(g-1)^{-1}$.

Заключение

В статье рассмотрены основные этапы построения семантического поля в некоторой проблемно-ориентированной предметной области. Описаны этапы метода анализа связей, основные показатели семантической сети, вычисление которых позволяют эксперту анализировать данные сети.

Литература

1. *Klerks P.* (2001). The network paradigm applied to criminal organizations: Theoretical nitpicking or a relevant doctrine for investigators? Recent developments in the Netherlands». Connections 24: 53–65. CiteSeerX: 10.1.1.129.4720.

2. *Спиридонов В.Ф.* Психология мышления: Решение задач и проблем: учебное пособие. М.: Генезис, 2006. 319 с. (Сер. «Учебник XXI века»).

3. *Kozma L.* K-algorithm Nearest Neighbours A, Helsinki University of Technology, 2008. URL: http://www.lkozma.net/knn2.pdf

4. *Саати Т.Л.* Принятие решений при зависимостях и обратных связях. Аналитические сети М.: Издательство ЛКИ, 2008. 360 с.

5. *Freeman L.* Centrality in social networks, conceptual clarifications // Soc. Networks. 1979. Vol. 1. P. 215–236.

Model representation of the mechanisms of formation of problem-oriented semantic fields

Alexander Sergeevich Sigov*, academician of RAS, President, Moscow state technical University Radioengineering, electronics and automation*

Valentin Viktorovich Nechaev*, academician of RANS, head of DEP, Moscow state technical University Radioengineering, electronics and automation*

Trofimenko Vsevolod Mikhailovich*, PhD student, Moscow state technical University Radioengineering, electronics and automation*

Describes one of possible approaches to the problem of constructing the problem-oriented semantic field in a problem-oriented areas of expertise through the provision of lexical units of unstructured or semi-structured text documents. The basis of constructing the model of the semantic field laid some of the scientific facts from cognitive psychology and psycholinguistics. The solution of the problem is carried out through the allocation of lexical units of semi-text documents.

Keywords – the semantic field, bionic approach, data mining, heuristic methods, heuristic modeling