Решение некоторых задач Text Mining при помощи концептуальных графов

© М.Ю. Богатырев, В.В. Тюхтин

Тульский государственный университет
okkambo@mail.ru

Аннотация

В работе задача извлечения ассоциативных правил из текстовых данных сформулирована на концептуальных графах – семантических моделях текстов. Показано, что способом решения данной и других задач Text Mining является кластеризация концептуальных графов. Данный подход применен в проекте экспериментальной электронной библиотеки научных статей.

1 Введение

Известной тенденцией развития электронных библиотек является их превращение в *системы поддержки знаний*. Поддержка знаний означает решение на данных библиотеки специфических задач обработки текстов с целью построения объектов, трактуемых как *знания* в рамках определенных моделей знаний.

Методы, модели и алгоритмы, применяемые в подобных задачах обработки текстов, образуют направление, известное как Text Mining [11].

Это направление связано с другим известным направлением, Data Mining – извлечением знаний из данных [3]. Справедливо считать, что Text Mining – это Data Mining применительно к текстовым данным. До сих пор нет устоявшегося русского термина – перевода для Text Mining, поэтому мы будем употреблять англоязычный термин.

Среди методов Text Mining можно выделить методы, применяемые непосредственно к текстам как неструктурированным данным, и методы, использующие структурированные модели текстов.

Одной из востребованных здесь структурированных моделей текста является *концептуальный граф* [10]. Вместе с *концептуальными решетками* концептуальные графы относятся к *концептуальным структурам*, которые являются одним из формальных представлений знаний. Концептуальные графы как атомарные объекты моделирования, в отличие, например, от ключевых слов текста, предоставляют информацию о структуре предложения, его логике и семантике, что важно в задачах Text Mining.

Концептуальный граф – это двудольный направленный граф, состоящий из двух типов узлов: *концептов* и *концептуальных отношений*. Концептуальный граф соответствует одному предложению текста. Концептами концептуального графа являются части речи – существительные, прилагательные, глаголы, наречия. Концептуальные отношения не являются элементами текста непосредственно, но строятся, например, на основе анализа семантических ролей предложения [5].

В работах [13–15] рассмотрена методика применения концептуальных графов в качестве объектов хранения электронных библиотек, рассмотрен специальный, эволюционный подход к задачам кластеризации концептуальных графов. В данной работе концептуальные графы применяются к решению некоторых задач Text Mining.

2 Задачи Text Mining

В задачах Text Mining строятся *кластеры*, *ассоциации*, анализируются *особенности* текстов, *подобие* текстов и т.д.[4], [9]. Традиционные методы обработки текстов используют ключевые слова или векторные модели текста, что требует значительных вычислительных ресурсов при обработке больших текстовых коллекций.

Концептуальный граф как модель сложнее, чем набор ключевых слов, но компактнее, чем вектор, построенный на тексте, например, в методах латентно-семантического анализа [5]. Это позволяет эффективно применять концептуальные графы в задачах Text Mining.

Применение концептуальных графов привносит определенную специфику в задачи Text Mining. Далее рассмотрены две задачи – построение ассоциативных правил и поиск особенностей в тексте, решения которых основаны на решении задачи кластеризации концептуальных графов.

Всякая технология Text Mining включает в себя два этапа:

- подготовку данных;
- решение задач Text Mining на подготовленных данных.

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL'2008, Дубна, Россия, 2008.

Рассмотрим реализации указанных этапов с применением концептуальных графов.

2.1 Подготовка данных

Применительно к текстам, подготовка данных означает замену неструктурированных текстов их структурированными моделями. В данном случае текстам должны быть сопоставлены концептуальные графы.

Проблема построения концептуальных графов хорошо известна [2], [6], [9] и не имеет общего решения, ввиду семантической неоднозначности концептуальных графов.

Следующий хрестоматийный пример демонстрирует такую неоднозначность: если спросить, о чем повествует предложение «*Студент читает книгу*»?, то ответами могут быть «о студенте», «о чтении», «о книге». Поэтому возможны три варианта концептуальных графов Направления дуг графа

[Студент] ← (Агенс) ← [Читать] →
→ (Пациенс) → [Книга] (1)

выделяют концепт [Читать] как основной, поэтому данный граф определяет, что рассматриваемое предложение повествует, прежде всего, о чтении.

Здесь использован *вербоцентрический* подход к построению концептуальных графов и линейная форма их представления [10]: в прямоугольные скобки заключаются концепты, в круглые – отношения.

На практике применяют ручное и автоматическое построение концептуальных графов. Ручное построение разрешает возможные семантические неоднозначности и считается наиболее корректным. Однако, для создания полнофункциональных систем на основе концептуальных графов, особенно систем, работающих в реальном времени, принципиальное значение имеет автоматическое построение графов.

Известны подобные англоязычные системы [2], [6]. Нами был разработан и опробован алгоритм построения концептуальных графов для предложений русского языка [15]. Для реализации алгоритма необходима подсистема синтаксического разбора предложений, например, система DWARF [17].

2.2 Построение ассоциативных правил

Построение ассоциативных правил является классической задачей Data Mining и заключается в следующем.

Пусть $R = \{r_1, ..., r_n\}$ – множество объектов, $T = \{t_1, ..., t_m\}$ – множество транзакций, такое, что каждая транзакция есть эксперимент по обнаружению объектов из R и представляет собой бинарный вектор длины n, в котором i-я компонента равна 1, если объект r_i присутствует в транзакции. Пусть $X \subset R, Y \subset R \setminus X$ – подмножества объектов в множестве R.

Ассоциативное правило на R есть клауза:

$$X \Rightarrow Y \quad (2)$$

Правило (2) поддерживается с *доверием* γ, если

$$\frac{\sum_{j=1}^{m} t_j[X \wedge Y]}{\sum_{j=1}^{m} t_j[X]} \geq \gamma \quad (3)$$

и имеет *распространенность* σ, если

$$\frac{\sum_{j=1}^{m} t_j[X \wedge Y]}{m} \geq \sigma \quad (4)$$

Здесь $t_j[X \wedge Y]$ – транзакция, в которой обнаружены объекты X и объекты Y, $t_j[X]$ - транзакция, в которой обнаружены объекты X.

Неравенство (3) обычно трактуется как вероятность того, что появление объектов X влечет появление объектов Y, что вполне вписывается в статистический подход к построению ассоциативных правил [3]. Правила (2) также часто интерпретируют как правила типа «если – то», не учитывая, что $\Rightarrow$ - знак метаимпликации.

Логическая интерпретация правил (2) как клауз применяется реже, но при использовании концептуальных графов эта интерпретация – «от общего к частному» - как раз актуальна.

Ассоциативное правило на концептуальных графах определяется следующим образом. Пусть $G = \{g_1, ..., g_n\}$ – множество концептуальных графов. *Обобщенным графом* для графа g назовем граф $\hat{g}$, полученный из графа g применением к нему конечного числа операций *разъединения* и *обобщения* [10]. Операции разъединения применяются к дугам графа, а операции обобщения – к его концептам и отношениям.

Граф (1) является обобщенным графом для графа, который соответствует, например, предложению «*Студент увлеченно читает интересную книгу*».

Обобщенный граф может быть порожден не одним графом, а множеством графов $\{g_i\}$.

Ассоциативным правилом на концептуальных графах назовем клаузу

$$\hat{g} \Rightarrow g_i(\text{б}, \text{в}), \quad (5)$$

где б, в – доверие и распространенность правила, вычисляемые следующим образом:

$$\frac{N[\hat{g}, g]}{N[\hat{g}]} \geq \alpha, \quad (6)$$

$$\frac{N[g]}{n} \geq \beta \quad (7)$$

Здесь $N[\hat{g}, g]$ - функция, возвращающая число графов из G, имеющих графы $\hat{g}$ и g в качестве под-

графов, аналогично интерпретируются $N[\hat{g}]$ и $N[g]$.

Множество транзакций как результат экспериментов по обнаружению объектов в данном случае отсутствует. Роль таких экспериментов играет операция обобщения графов.

Семантическая интерпретация правила (5) состоит в упомянутом уже принципе «от общего к частному»: обобщенный граф $\hat{g}$ задает общий смысл, графы g_i(б, в) – детали. Соответственно, один обобщенный граф может порождать несколько правил с различными величинами доверия и распространенности.

Определив ассоциативные правила в виде (5), мы сталкиваемся с проблемой их построения.

Покажем, что решение данной проблемы может быть найдено как решение задачи кластеризации.

В определении обобщенного графа не указан принцип, в соответствии с которым выполняются операции разъединения и обобщения. Эти операции, в действительности, служат моделью выделения на множестве G подграфов, обладающих некоторыми общими свойствами.

Такие свойства можно задать бинарным отношением близости графов в соответствие с некоторой мерой.

Пусть

$$C_1 \subset C_2 \subset ... \subset C_k \qquad (8)$$

– иерархия кластеров на множестве концептуальных графов, построенная с использованием некоторой меры близости графов.

Обозначим $g_l^{(i)} \in C_i$ l-й граф, принадлежащий кластеру C_i, $\hat{g}_s^{(i)}$ – его обобщенный граф.

Утверждение 1. Наличие кластеризации (8) достаточно для того, чтобы среди множества ассоциативных правил, существующих на множестве концептуальных графов, существовали правила вида

$$\hat{g}_s^{(i)} \Rightarrow g_r^{(j)}(\alpha, \beta) \qquad (9)$$

где $i > j$, индексы s, r не обязательно различны. Другими словами:

• кластеризация концептуальных графов позволяет получать обобщения концептуальных графов;

• иерархия кластеров на множестве концептуальных графов задает некоторое подмножество ассоциативных правил.

Решение задачи кластеризации определяется применяемой мерой близости.

В работе [13] рассматриваются меры близости концептуальных графов на основе известных коэффициентов Дайса – концептуальной близости d_c и относительной близости d_r:

$$d_c = \frac{2n(g_c)}{n(g_i) + n(g_j)}, \qquad (10)$$

$$d_r = \frac{2m(g_c)}{m_{g_c}(g_i) + m_{g_c}(g_j)}, \qquad (11)$$

где $g_c = g_i \cap g_j$, $n(g_i)$ – число концептов графа g_i, $m(g_c)$ – число отношений концептуального графа g_c, $m_{g_c}(g_i)$ – число отношений концептуального графа g_i, для которых хотя бы одна из вершин принадлежит графу g_c. Там же рассмотрены модификации указанных мер, учитывающих размеры сравниваемых графов, и наличие общезначимых концептов.

Применяя меры (6), (7), можно выделить на множестве G кластеры, графы в которых близки в смысле наличия у них одного обобщенного графа. Такой кластеризации можно сопоставить операцию разъединения.

Операция обобщения сложнее. Ее результатом может быть замена концептов и отношений на «более общие». Это требует введения *иерархии типов* концептов $\mathbf{T}_C$ и отношений $\mathbf{T}_R$.

Вершина каждой иерархии представлена *универсальным концептом* и *универсальным отношением* [10] в контексте решаемой задачи. Соответственно, внизу иерархий находятся типы *абсурдного концепта* и *абсурдного отношения*. Место типа в иерархии определяется *порядком* типа. Отношения в иерархии типов отношений характеризуются также *арностью отношения*.

В следующем разделе приведены некоторые примеры иерархий типов.

Наличие иерархий позволяет применить иные меры близости, измеряющие близость двух концептов как сумму расстояний от них до другого концепта с более высоким порядком в иерархии типов. Концепты разных иерархий типов могут быть связаны отношениями в конкретных графах.

Данная идея не нова [8], ее применение позволяет модифицировать меру (10) следующим образом.

Если два графа не имеют общих концептов, но в них есть концепты, принадлежащие одной иерархии типов, то эти концепты «засчитываются» в меру близости:

$$d_c = \frac{2n(g_c) + a(\sum_k f(c_{k,i}, c^0) + \sum_l f(c_{l,j} c^0))}{n(g_i) + n(g_j)}, \qquad (12)$$

где $c_{k,i}$, $c_{l,j}$ – k-й и l-й концепты, соответственно, i-го и j-го графа, $f(c_{k,i}, c^0)$ – функция расстояния между концептами графов и ближайшим общим концептом в иерархии, a – масштабирующий множитель. В простейшем варианте функция $f(c_{k,i}, c^0)$ возвращает натуральное число – разницу в уровнях иерархии между концептами c^0 и $c_{k,i}$.

Поддержка иерархии типов отношений не столь актуальна, ввиду того отношение по своей природе

семантически локально. В самом деле, например, отношения *агенс*, *пациенс*, *атрибут*, *цель* и т.п. действуют в рамках одного предложения и не являются наследниками каких-то более общих отношений.

2.3 Выделение особенностей в тексте

Плоская кластеризация множества концептуальных графов разбивает его на подмножества, слабо связанные с точки зрения применяемой меры близости графов. Поэтому плоская кластеризация может служить средством выделения тем (особенностей) в тексте. Данный подход известен [7], однако мы воспользуемся более общим подходом, использующим понятие регулярности [1].

Назовем концептуальный граф *регулярным*, если он является обобщенным графом хотя бы для одного другого графа. На практике свойство регулярности на концептуальных графах поддерживается множествами графов, имеющих обобщенные графы.

Назовем концептуальный граф *особенным*, если он не имеет ни одного регулярного графа.

Утверждение 2. Пусть G – множество концептуальных графов. Если имеет место кластеризация вида

$$G = \{C_1 \subset C_2 \subset ... \subset C_m, C_n\} \qquad (13)$$

то в кластере Cn существуют особенные графы.

Таким образом, особенный кластер, не входящий в иерархию (8), содержит особенные графы.

Множество особенных графов на множестве G соответствует фрагментам текста, имеющим особенности.

Поиск таких фрагментов, выпадающих из общей содержательной картины текста, представляет практический интерес.

Утверждения 1, 2 следует принимать с определенными оговорками. Оба утверждения являются достаточными, то есть не гарантируют необходимость существования декларируемых ассоциативных правил и особенных графов. Тем не менее, утверждения 1, 2 важны тем, что они условно связывают язык концептуальных графов, элементы которого здесь рассмотрены, с операцией кластеризации.

Таким образом, кластеризация становится экспериментальным инструментом решения рассмотренных задач Text Mining.

3 Проект электронной библиотеки

Рассмотренные принципы и решения применены в экспериментальном проекте электронной библиотеки научных статей.

Вместе с реляционной базой данных, содержащей тексты и их параметры, в системе поддерживается XML - база данных концептуальных графов. Выбор данного формата базы обусловлен тем, что языки концептуальных графов [10], [13] используют XML.

В системе возможны ручное и автоматическое построение концептуальных графов. Концептуальные графы строятся только для аннотаций статей, что позволяет иметь достаточно компактные графы. Аннотации статей загружаются с сервера [18]. Также выполняется обработка электронного ресурса статей конференция RCDL [19].

В настоящее время в системе реализованы следующие функции:

- полнотекстовый поиск в статей;
- построение концептуальных графов;
- кластеризация концептуальных графов;
- визуализация результатов кластеризации в виде дендрограмм (актуальна только в экспериментах с небольшим числом характерных графов).

Приведем некоторые характерные результаты экспериментального решения рассмотренных задач.

В экспериментах исследовались возможности эволюционного подхода к решению задачи кластеризации на концептуальных графах.

Эволюционный подход, описанный в работах [13, 14, 16], основан на применении генетических алгоритмов оптимизации при поиске экстремума целевой функции, отражающей меру близости концептуальных графов друг другу. Использовались меры близости (10) – (12).

Исследовались варианты кодирования популяций в схемах генетических алгоритмов, алгоритмы рекомбинации, влияние вероятностных параметров алгоритма на сходимость.

Моделирование семантики аннотаций текстов научных статей при помощи концептуальных графов позволило обнаружить особые типы аннотаций, важные как средства репрезентативности корпуса текстов научной тематики – это аннотации *декларативного* типа и аннотации *повествовательного* типа.

Аннотация декларативного типа содержит небольшое число предложений. Концептуальная близость предложений невелика. Относительная близость может быть высокой за счет применения типичных оборотов речи, характерных для подобного типа аннотаций.

Аннотация повествовательного типа содержит большее число предложений, чем аннотация декларативного типа. Концептуальная близость предложений высокая или фрагментарно высокая, когда концептуально связаны несколько предложений, составляющих фрагмент текста.

В эксперименте, результаты которого представлены на рис. 1, 2 выбрана длинная повествовательная аннотация из 11 предложений (пронумерованных от 0 до 10):

0. We assume that the modality (i.e., number of local optima) of a fitness landscape is related to the difficulty of finding the best point on that landscape by evolutionary computation (e.g., hillclimbers and genetic algorithms (GAs)).
1. We first examine the limits of modality by constructing a unimodal function and a maximally multimodal function.

2. At such extremes our intuition breaks down.
3. A fitness landscape consisting entirely of a single hill leading to the global optimum proves to be hard for hillclimbers but apparently easy for GAs.
4. A provably maximally multimodal function, in which half the points in the search space are local optima, can be easy for both hillclimbers and GAs.
5. Exploring the more realistic intermediate range between the extremes of modality, we construct local optima with varying degrees of “attraction” to our evolutionary algorithms.
6. Most work on optima and their basins of attraction has focused on hills and hillclimbers, while some research has explored attraction for the GA's crossover operator.
7. We extend the latter results by defining and implementing maximal partial deception in problems with k arbitrarily placed global optima.
8. This allows us to create functions with multiple local optima attractive to crossover.
9. The resulting maximally deceptive function has several local optima, in addition to the global optima, each with various size basins of attraction for hillclimbers as well as attraction for GA crossover.
10. This minimum distance function seems to be a powerful new tool for generalizing deception and relating hillclimbers (and Hamming space) to GAs and crossover. Все предложения аннотации логически связаны.

Другая аннотация состоит из одного предложения:

11. This paper describes an initial version of a library of sharable and reusable medical ontological theories, organized according to a proposed classification of ontologies.

Длинная повествовательная аннотация содержит большое число различных слов, которые могут также встречаться во второй короткой аннотации.

Тем не менее, необходимо отделить единственное, последнее по номеру предложение 11 от остальных.

Обе меры близости, использованные в экспериментах рис. 1, 2, обеспечивают успешную кластеризацию: граф с номером 11 представляет собой отдельный кластер. Из рис. 2 видна иерархичность полученной кластеризации и соответствие ее виду (13). Это следствие «*повествовательности*» первой аннотации из 11 предложений, проявляющейся во «*вложенности смыслов*» предложений. Соответственно, на полученных кластерах имеют место ассоциативные правила вида (5). Граф с номером 11 – особенный. Ему соответствует предложение, образующее *особенную аннотацию*.

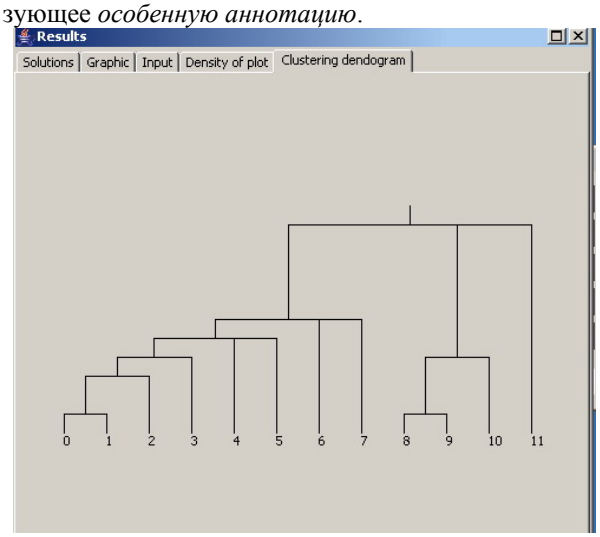

Рис. 1. Результат кластеризации генетическим алгоритмом с использованием концептуальной меры близости в виде дендрограммы (внизу показаны номера концептуальных графов)

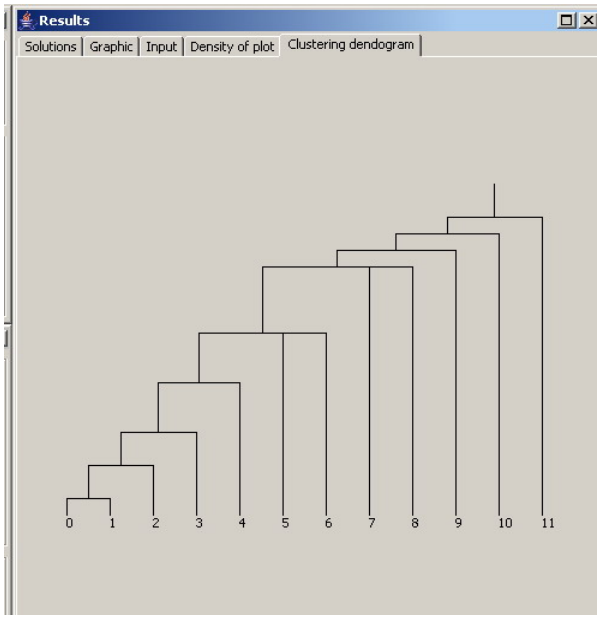

Рис. 2. Результат кластеризации генетическим алгоритмом с использованием реляционной меры близости

4 Выводы и дальнейшие исследования

Таким образом, применяя кластеризацию концептуальных графов, можно решать некоторые задачи Text Mining на текстах, моделируемых при помощи концептуальных графов. При этом реляционная мера близости концептуальных графов полнее отображает семантическую близость текстов.

Другим, интересным инструментов анализа концептуальных графов является аппарат математиче-

ской логики, предложенный еще в пионерских работах Р. Совы [10]

Применение данного аппарат позволит исследовать ассоциативные правила как клаузы, что следует из их строгого определения (5)

Это позволит получать математически строгие результаты, имеющие новые приложения

Литература

[1] Agrawal, R., A. Arning, T. Bollinger, M. Mehta, J. Shafer, R. Srikant. The Quest Data Mining System, Proc. of the 2nd International Conference on Knowledge Discovery in Databases and Data Mining, Portland, Oregon, August, 1996.

[2] Boytcheva, S. Dobrev, P. Angelova, G.CGExtract: Towards Extraction of Conceptual Graphs from Controlled English. Lecture Notes in Computer Science № 2120, Springer 2001.

[3] D. Hand, H. Mannila, P. Smyth Principles of Data Mining. MIT Press, Cambridge, MA, 2001, 445 p.

[4] Feldman D., Hirsh M., Mining Associations in Text in the Presence of Background Knowledge.- Proc. of the 2nd International Conference on Knowledge Discovery (KDD-96), Portland, 1996.

[5] Gildea, D., Jurafsky, D.: Automatic labeling of semantic roles. Computational Linguistics 28(3), 2002, p.p. 245--288.

[6] Hensman, S., Dunnion, J. Applying Verbnet For Automatic Semantic Role Identification. Обработка текста и когнитивные технологии : Сб. / Под ред. Полякова В.Н., Соловьева В.Д. – Варна: 2003 (Вып.8), с. 471–481.

[7] Knorr, N. Algorithms for Mining Distance-based Outliers in Large Datasets, Proc. of the International Conference on Very Large Data Bases (VLDB'98), Newport Beach, CA, 1998.

[8] Foo, N., Garner, B. J., Rao, A., Tsui, E. Semantic distance in conceptual graphs. In: Conceptual structures: current research and practice book contents. Ellis Horwood Upper Saddle River, NJ, USA, 1992. P. 149–154.

[9] Montes-y-Gomez, M., Gelbukh, A., Lopez-Lopez, M. Text Mining at Detail Level Using Conceptual Graphs. Lecture Notes In Computer Science; Vol. 2393. P. 122–136.

[10] Sowa R., Conceptual Graphs: Draft Proposed American National Standard, International Conference on Conceptual Structures ICCS-99, Lecture Notes in Artificial Intelligence 1640, Springer 1999.

[11] Tan, A. Text Mining: The state of the art and challenges, Proc. of the Workshop Knowledge Discovery from advanced Databases PAKDDD-99, April 1999.

[12] Wermelinger M. Conceptual Graphs and First Order Logic. - 3 rd International Conference on Conceptual Structures. Lecture Notes in Artificial Intelligence, Vol 954. – Springer Verlag, 1995, p. 323–337.

[13] Богатырев М.Ю., Латов В.Е., Столбовская И.А. Применение концептуальных графов в системах поддержки электронных библиотек. – Электронные библиотеки: перспективные методы и технологии, электронные коллекции // Труды Девятой Всероссийской научной конференции RCDL'2007 (Переславль-Залесский, Россия, 15–18 октября 2007). – Т. 2, С. 104–110.

[14] Богатырев М.Ю., Латов В.Е., Столбовская И.А., Тюхтин В.В. Эволюционный подход к задаче кластеризации на концептуальных графах и его применение в системах поддержки электронных библиотек. – Математические методы распознавания образов. Тринадцатая Всероссийская конференция : Сб. докладов. – М. : МАКС Пресс, 2007. – 668 с. – С. 464–468.

[15] Богатырев М.Ю., Тюхтин В.В. Построение и анализ концептуальных графов в системах поддержки электронных библиотек. – Известия ТулГУ. Сер. Технические науки, 2008 (в печати).

[16] Богатырёв М.Ю., Латов В.Е. Исследование генетических алгоритмов кластеризации. – Известия. ТулГУ. Сер. Математика Механика. Информатика. Том 8, вып. 3 . Информатика. – Тула, 2002. – С. 101–107.

[17] Электронный ресурс: Cognitive Technologies – Интеллектуальные технологии управления. http://www.cognitive.ru

[18] Электронный ресурс: Scientific Literature Digital Library http://citeseer.ist.psu.edu/

[19] Электронный ресурс: Труды конференций Электронные библиотеки: Перспективные Методы и Технологии, Электронные коллекции http://rcdl.ru/

Solving Some Text Mining Problems with Conceptual Graphs

M.Y. Bogatyrev, V.V. Tuhtin

The problem of finding Associated Rules and other Text Minin problems formulated on Conceptual Graphs. It is shown that Clustering is appropriate way to solve these problems.