Эволюционный подход к задаче кластеризации на концептуальных графах и его применение в системах поддержки электронных библиотек

Богатырев М. Ю., Латов В. Е., Столбовская И. А., Тюхтин В. В.
okkambo@mail.ru
Тульский государственный университет

Одним из направлений развития современных электронных библиотек является расширение их функциональных возможностей. Такое расширение было бы значительным, если бы в библиотеках хранились не только тексты, но и их смысловое содержание. Решить эту задачу можно, исследуя и реализуя семантические модели текстов.

В данной работе предлагается применить одну из семантических моделей текста — *концептуальный граф* [1] — в качестве объекта хранения электронной библиотеки.

Применение концептуальных графов позволяет развивать технологии поддержки электронных библиотек, по крайней мере, в двух направлениях:

1) автоматизация построения каталогов библиотек, модификация и коррекция существующих каталогов на основе анализа потока входных текстов;
2) извлечение знаний из электронных библиотек в виде *концепций* и *онтологий*.

Оба указанных направления в настоящее время представлены множеством методов и технологий [2, 3]. Эффективное применение здесь концептуальных графов связано с решением задач *агрегирования* и *кластеризации* на графах.

В работе содержится обзор современных подходов к использованию концептуальных графов и некоторые результаты авторов, касающиеся решения задач кластеризации на концептуальных графах.

Концептуальные графы и их применение

Вместе с *концептуальными решетками* концептуальные графы относятся к *концептуальным структурам*, которые являются одним из формальных представлений знаний [4].

Концептуальный граф — это двудольный направленный граф, состоящий из двух типов узлов: *концептов* и *концептуальных отношений*.

Разработан стандарт представления концептуальных графов [1] и языки их описания, среди которых наиболее популярны CGIF (Conceptual Graph Interchange Form) и XML-представление концептуальных графов.

В системах поддержки электронных библиотек важными задачами являются задачи автоматической классификации и автоматической фильтрации входных документов при известном множестве тематических интересов. Данное множество может быть представлено как система концепций (концептов), выражаемая через концептуальные графы.

В результате, задачи классификации сводятся к решению фундаментальной задачи кластеризации на графах.

Кластеризация на концептуальных графах

В любой задаче кластеризации важной проблемой является построение меры близости кластеризуемых объектов. Под мерой близости концептуальных графов понимается количественная характеристика, призванная отразить семантическую близость порождающих графы предложений, что сделать, очевидно, полностью невозможно. Поэтому проблема близости остается центральной в анализе концептуальных графов.

Для двух графов G_1 и G_2 мера близости зависит от двух значений: концептуальной близости s_c и относительной близости s_r.

$$s_c = \frac{2n(G_c)}{n(G_1) + n(G_2)}, \tag{1}$$

где $G_c = G_1 \cap G_2$, $n(G)$ — число концепций — концептуальных узлов графа G.

$$s_r = \frac{2m(G_c)}{m_{G_c}(G_1) + m_{G_c}(G_2)}, \tag{2}$$

где m_{G_c} — число отношений — относительных узлов концептуального графа G_c, $m_{G_c}(G)$ — число отношений — относительных узлов концептуального графа G, для которых хотя бы одна из вершин принадлежит графу G_c.

Анализ мер близости (1), (2), выполненный в работе, демонстрирует их несовершенство. В результате применения мер близости (1), (2) близкими могут оказаться графы, имеющие просто большое число одинаковых концептов или отношений, но по смыслу совершенно далекие. Поэтому вместо мер близости (1), (2) предлагается использовать их модификации:

$$s_c = \frac{2n(G_c)l}{n(G_1) + n(G_2)}, \tag{3}$$

где

$$l = \begin{cases} k\frac{n(G_1)}{n(G_2)}, & \text{если } n(G_1) \geqslant n(G_2); \\ k\frac{n(G_2)}{n(G_1)}, & \text{если } n(G_1) < n(G_2); \end{cases}$$

k — масштабирующий коэффициент.

Для формулы (2):

$$m_{G_c}(G) = m_{\text{both}} + b_1 + b_2 + \cdots + b_i, \quad i = 1, \ldots, m - m_{\text{both}},$$

$b_i \in \{0, 1\}$, где m — число всех отношений графа G, m_{both} — число всех отношений графа G, для которых обе вершины принадлежит графу G_c, b_i — коэффициент общезначимости — принимает значения от 0 до 1, в зависимости от типа отношения.

Под мерой близости двух концептуальных графов, принимающей значение от 0 до 1, будем понимать значение

$$s = d_1 s_c + d_2 s_r, \tag{4}$$

где d_i — масштабирующие коэффициенты, определяемые экспериментально.

Как следует из экспериментов, данная мера близости повышает качество кластеризации, но не решает проблему в целом.

Эволюционный подход к кластеризации концептуальных графов

Изменение коэффициентов k и b_i, введенных выше, влечет появление множественных вариантов кластеризации даже в рамках одной меры близости концептуальных графов.

В работе предполагается исследование мер близости на концептуальных графах в рамках эволюционного подхода к решению задачи кластеризации, который состоит в применении эволюционных алгоритмов оптимизации при поиске экстремума целевой функции, отражающей меру близости графов друг другу.

Коэффициенты k и b_i в этом случае входят в целевую функцию в качестве параметров.

Эволюционные алгоритмы основаны на генетическом алгоритме.

1. Генетический алгоритм работает с множеством (популяцией) приближений решения задачи. В результате оптимальное решение может быть найдено как множество различных объектов, что характерно для задач со сложными целевыми функциями.
2. Генетический алгоритм эффективен при поиске экстремума целевых функций, имеющих конечные разрывы, что имеет место для функций, являющихся мерами близости концептуальных графов.

Известны несколько подходов к решению задач кластеризации с применением генетических алгоритмов [6, 7]. Реализация эволюционного подхода требует настройки параметров алгоритма: выбора системы кодирования, способа рекомбинации хромосом, введения механизма мутации [8].

В работе исследованы варианты настроек параметров генетического алгоритма классификации. В частности, построена специфическая кодировка решений, дающая высокое качество кластеризации. Кодировка использует хромосомы вида $a_1 a_2 \dots a_n$, где $a_i \in \{1, \dots, n\}$, n — количество объектов кластеризации.

В докладе представлены результаты вычислительных экспериментов на конкретном материале текстов аннотаций научных статей. Обсуждаются вопросы реализации подхода в информационной системе — пилотном проекте электронной библиотеки научных статей.

Работа выполнена при поддержке РФФИ, проект № 07-07-00276-а.

Литература

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

[2] *Городецкий В. И., Самойлов В. В., Малов А. О.* Современное состояние технологии извлечения знаний из баз и хранилищ данных // Журнал Российской ассоц. искусственного интеллекта. — 2002. — № 3. — С. 3–31.

[3] *Hirst G.* Ontology and the Lexicon // Handbook on Ontologies in Information Systems, Berlin: Springer, 2003.

[4] *Sarbo J.* Formal conceptual structure in language. In Dubois, D. M., editor, Proceedings of Computing Anticipatory Systems (CASYS98).— Woodbury, New York, 1999.— Pp. 289–300.

[5] *Montes-y-Gomez, Gelbukh, Lopez-Lopez, Baeza-Yates* Flexible Comparison of Conceptual Graphs // Lecture Notes in Computer Science 2113, Springer-Verlag, 2001.

[6] *Maulik U., Bandyopadhyay S.* Genetic algorithm-based clustering technique // Pattern Recognition. — 2000. — vol. 33. — Pp. 1455–1465.

[7] *Kivijarvi Juha, Lehtinen Joonas , Nevalainen Olli* A Parallel Genetic Algorithm for Clustering // Turku Centre for Computer Science, Tech. report № 469, August 2002.

[8] *Богатырёв М. Ю., Латов В. Е.* Исследование генетических алгоритмов кластеризации // Изв. ТулГУ. Сер. Математика. Механика. Информатика., Тула, 2002. — Т. 8, вып. 3. Информатика.— С. 101–107.