ПОСТРОЕНИЕ И АНАЛИЗ СВЕРХБОЛЬШИХ СЕМАНТИЧЕСКИХ СЕТЕЙ

Т.С. Лебедева, Е.Ю. Костюченко

Томский государственный университет систем управления и радиоэлектроники,

Россия, г. Томск, пр. Ленина, 40, 634050

E-mail: key@keva.tusur.ru

CONSTRUCTION AND ANALYSIS OF SUPER SEMANTIC NETWORKS

T.S. Lebedeva, E.Y. Kostyuchenko

Tomsk State University of Control Systems and Radioelectronics,

Russia, Tomsk, Lenin str., 40, 634050

E-mail: key@keva.tusur.ru

Abstract. *Within the framework of the work, an algorithm for constructing super-large semantic networks has been developed. It was tested using the Internet encyclopedia "Wikipedia" as an example. During the implementation, a parser for the syntax analysis of Wikipedia pages was developed. It finds the links between the articles. A graph based on list of articles and categories was formed. On the basis of the obtained graph analysis, algorithms for finding domains of high connectivity in a graph were proposed and tested.*

Введение. В самом конце XX века Альберта Барабаши и Реки Альберт опубликовали в своей статье модель сетевой структуры, которая формируется по правилу предпочтительного связывания. Модель представляла собой выращенный из графа-затравки случайный динамический граф. Это дало начало анализу социальных сетей в целом [1–4]. Данная модель, именуемая безмасштабная сеть (англ. *scale-free network*) в достаточной мере способна отобразить социальные, коммуникационные, биологические системы, а также графы цитирования [4].

Данная работа нацелена на создание методики поиска областей высокой связности в семантическом графе, получаемом при анализе такой сети. Тестирование проводится на электронной энциклопедии Википедия. Каждая статья, раскрывая описываемое в ней понятие, ссылается на соответствующие статьи согласно семантическим потребностям в интерпретации этого понятия. Задевая новые, смежные, косвенно или прямо зависимые понятия, создается сеть неизученных связей между областями знаний.

Поиск областей высокой связности. Выбор способа хранения и сбор данных более подробно описаны в [5]. В ходе исследований был реализован предложенный алгоритм работы с полученным семантическим графом (рис. 1). Основу алгоритма составляют два цикла – на добавление вершин и на удаление. Это позволяет оперативно контролировать состав предполагаемой предметной области, добавляя в нее новые вершины и удаляя вершины, переставшие удовлетворять критерию вхождения. Добавление вершин идет постепенно, с каждой новой добавленной вершиной параметры связности с предметной областью остальных вершин, находящихся в предметной области или за ее пределами, меняются.

Рассматривается несколько вариантов критерия включения вершины в предметную область.

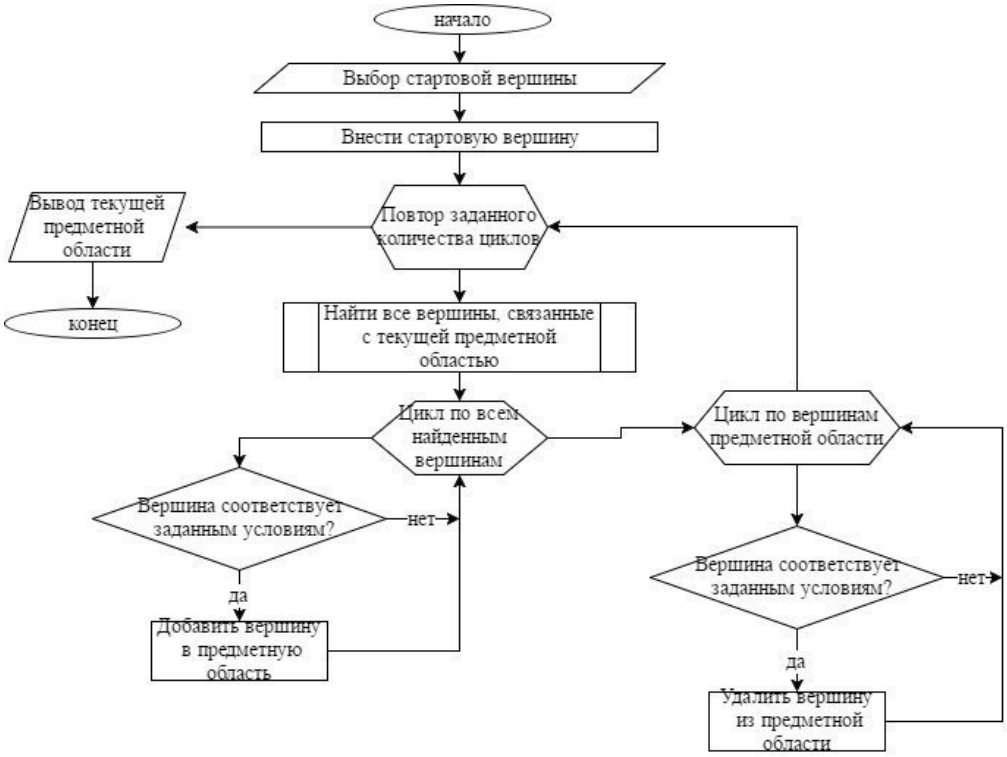

Рис. 1. Алгоритм нахождения областей высокой связности в предметной области

$$N = \frac{\text{количество связей вершины с ПО}}{\text{количество связей в ПО}};\quad K = \frac{\text{количество связей вершины с ПО}}{\text{количество всех связей вершины}}$$

N показывает, какой вклад связей вершина вносит в предметную область. Чем больше таких вершин в предметной области, тем сильнее связь внутри нее и тем быстрее она стабилизируется. При проверке проводится сравнение значения рассчитанного критерия для рассматриваемой вершины с пороговым значением. При превышении вершина добавляется (оставляется) в предметной области, в противном случае не добавляется (удаляется). Параметр *K* отвечает за значимость вклада для самой вершины.

Однако, тестирование данных двух вариантов критерия выявило следующие недостатки: вершины, имеющие мало связей, не могут попасть в сформированную по первому критерию предметную область, даже если все связи ведут в предметную область. С другой стороны, с помощью первого критерия мы можем контролировать объем формируемой предметной области. Второй вариант алгоритма не сужает предметную область из-за непрерывно растущего знаменателя дроби, область растет безгранично. Предложена комбинация этих двух критериев (через «или» на добавление и через «и» на исключение вершин). Размер в зависимости от значений порогов *K* и *N* представлены в таблице 1. Из таблицы четко видна зависимость между размером предметной области и параметрами. Следующим шагом проводимых исследований станет проверка устойчивости сформированной предметной области в зависимости от стартовой вершины. Очевидно, что предметная область для двух вершин, входящих в нее должна слабо отличаться в зависимости от стартовой вершины. По предварительным итогам тестирования это не так. Проводится доработка предложенного критерия для обеспечения устойчивости формируемой предметной области.

Таблица 1

Зависимость размеров предметной области от пороговых значений, прочерк – область не стабилизируется

K \ *N*	0,01	0,02	0,05	0,1	0,2
0,33	-	-	-	-	-
0,5	-	123	49	63	28
0,53	-	96	46	24	12
0,7	128	84	45	24	9
0,9	114	85	45	24	9
1	114	85	45	24	9

Заключение. В ходе работы реализовано построение семантического графа с учетом большого числа связей между вершинами и оптимизации поиска вершин, смежных с рассматриваемыми. Разработан алгоритм построения предметной области. Экспериментально выявлены и теоретически обоснованы недостатки применения каждого из критериев вхождения по отдельности. Предложена комбинация этих критериев, позволяющих получать стабильную предметную область, проведено исследование зависимости ее размера от значений порогов. Выявлена проблема устойчивости формируемой предметной области в зависимости от стартовой вершина. Решению данной проблемы будет посвящен следующий этап исследования. Кроме того, в случае формирования эталонных вариантов предметных областей возможно применение метаэвристичесих методов, базирующихся, например, на глубоком обучении нейронных сетей [6].

Благодарности. Проект выполнен в рамках базовой части государственного задания министерства образования и науки Российской Федерации на 2017–2019 гг. Номер 8.9628.2017/БЧ.

СПИСОК ЛИТЕРАТУРЫ

1. Евин И.А. Введение в теорию сложных сетей / И.А. Евин // Компьютерных исследования и моделирование. –2010. – Т. 2. – № 2. – С. 121–141.
2. Barab´asi L.A. Emergence of scaling in random networks / L.A. Barab´asi, R. Albert // Science. – 1999. – V. 286. – P. 509–512.
3. Barab´asi L.A. Scale-free characteristics of random networks: the topology of the world-wide web / L.A. Barab´asi, R. Albert, H. Jeong // Physica A. – 2000. – V. 281. – P. 69–77.
4. Albert R. Diameter of the world-wide web/ R. Albert, H. Jeong, L.A. Barab´asi // Nature. – 1999. – V. 401. – P. 130–131.
5. Лебедева Т.С. Поиск предметных областей на основе большого семантического графа / Т.С. Лебедева, Е.Ю. Костюченко // Многоядерные процессоры, параллельное программирование, ПЛИС, системы обработки сигналов. – 2016. – №. 6. – С. 145–152.
6. Kipyatkova I.S. Variants of deep artificial neural networks for speech recognition systems / I.S. Kipyatkova, A.A. Karpov // SPIIRAS Proceedings. – 2016. – V. 49. – No. 6. – P. 80–103.