АНАЛИЗ СЕТЕВОЙ МОДЕЛИ ВИКИПЕДИИ

Макрушин С. В.
(*Финансовый университет при Правительстве Российской Федерации, Москва*)

Большое количество исследований теории сложных сетей (ТСС) посвящено анализу массивов гипертекстовых документов, среди которых большое внимание уделяется Википедии. Изучение структуры Википедии имеет большую практическую ценность, так как она обладает уникальным сочетанием полноты информации о мире и структурированности, а на ее основе можно строить семантические сети (онтологии), справочные системы, системы информационного поиска, системы семантического анализа текстов и приложения других типов [2].

В рамках исследования была обработана распространяемая фондом Викимедиа база данных русскоязычного сегмента Википедии, насчитывающая более 5,3 млн страниц и более 150 млн ссылок между ними. Для анализа были отобраны только страницы, являющиеся словарными статьями (далее – статьи) и категории (используются для тематического структурирования Википедии). В результате была получена сеть (граф) с характеристиками, приведенными в Таблице 1.

Таблица 1. Интегральные показатели сети Википедии, построенной в рамках исследования

Типы страниц	*Количество страниц (тыс. шт.)*	*Ссылок (тыс. шт.)*		*Удельное количество ссылок*	
		на статьи	*на категории*	*на статьи*	*на категории*
Статьи	3 305	92 167	9 187	27,9	2,8
Категории	405	9 187	770	*со статей:* 22,7	*на родит. кат-ии:* 1,9
Всего	**3 710**	**101 355**	**9 958**		

Для построенной сети был выполнен анализ распределения степеней узлов. Для повышения качества анализа хвоста распределения были использованы непараметрические методы оценки плотности распределения на основе ядерного сглаживания (kernel density estimation) и показана целесообразность применения методов выбора локально-адаптивной ширины окна сглаживания. Анализ позволил выявить, что плотность распределения исходящих степеней узлов для больших степеней узлов имеет значительные флуктуации (до двух порядков величины) из-за наличия в Википедии шаблонов (специальных элементов оформления статей), имеющих очень большое (до нескольких сотен) количество ссылок.

Анализ распределения входящих степеней узлов показал, что оно может хорошо описываться степенным законом распределения. Сети с таким распределением вершин в ТСС принято называть безмасштабными, а их формирование хорошо описывается модель Барбаши-Альберта [1]. Согласно этой модели, сеть возникает в результате пошагового роста, при котором действует принцип «предпочтительного присоединения», т.е. новые связи с большей вероятностью образуются у узлов, уже имеющих большое количество связей, что хорошо согласуется с логикой развития сети Википедии.

Очень существенно от степенного закона распределения отличается распределение исходящих степеней узлов сети. Это вызвано существенно иной природой возникновения исходящих ссылок в статьях: они формируются в процессе правки статьи и зависят от ее проработанности и целесообразности добавления новых ссылок, а не от результата большого количества правок на миллионах других статей. Таким образом, модель Барбаши-Альберта лишь частично применима для описания формирования структуры сети Википедии. Адаптация этой модели для сети Википедии является предметом дальнейших исследований авторов.

Литература

1. A-L BARABASI, R. ALBERT, *Emergence of scaling in random networks* // Science 10/1999 №286 (№5439): 509–512.