ФРАКТАЛЬНАЯ МОДЕЛЬ ИНФОРМАЦИОННОГО ПРОСТРАНСТВА

Кудряшова Э.Е., Копылова М.Ю., Чистов Д.А.

Волгоградский государственный технический университет

Волгоград, Россия

В работе проводится моделирование Web-пространства на основе фрактального подхода, базирующегося на свойстве самоподобия информационного пространства. Вычисляются основные характеристики Web-сайтов, такие как ранг сайта, количество ссылок на сайт, частота упоминания сайта, а также для каждого Web-сайта определяется коэффициент Ципфа на основе закона Ципфа.

Целью исследования является проектирование модели информационного пространства на базе фрактального подхода и вычисление емкости информационного пространства.

Работа включает в себя решение следующих задач: проведение анализа топологии информационного пространства; изучение прямой и обратной сетевой навигации на основе определенных Web-узлов; проектирование модели информационного пространства на основе фрактального подхода, базирующегося на свойстве самоподобия информационного пространства; получение константы Ципфа для исследуемых Web-узлов; вычисление емкости информационного пространства; обобщение полученных результатов.

В настоящее время существуют некоторые попытки изучения топологии информационного пространства, однако четкой теории предложено не было. Знание топологии информационного пространства позволяет реализовать концепцию сетевой навигации (как прямой, следуя гиперссылкам, так и обратной). Теория фракталов находит свои приложения в разных областях, в том числе и при анализе информационных потоков. Web-пространство, которое является динамичной частью информационного пространства, можно рассматривать как среду, характеризующуюся большим количеством скрытых в нем неявных экспертных оценок, реализованных в виде гиперссылок. Информационное пространство в целом, ввиду его объемов и динамики изменения, принято рассматривать как стохастическое.

Фрактальный подход базируется на свойстве самоподобия информационного пространства, то есть сохранение внутренней структуры множеств при изменениях их размеров или масштабов их рассмотрения извне. Самоподобие информационного пространства выражается, прежде всего, в том, что при почти обвальном росте этого пространства в последние десятилетия, гиперболические частотные и ранговые распределения, получаемые в таких разрезах, как источники и авторы, практически не меняют своей формы.

Дж. Ципф на основе статистического материала показал, что распределение слов естественного языка подчиняется простому закону, который можно сформулировать следующим образом: «Если к какому-либо достаточно большому тексту составить список всех встретившихся в нем слов, затем расположить эти слова в порядке убывания частоты их встречаемости в данном тексте и пронумеровать в порядке от 1 (порядковый номер наиболее часто встречающегося слова) до R, то для любого слова произведение его порядкового номера (ранга) в таком списке и частоты его встречаемости в тексте будет величиной постоянной, имеющей примерно одинаковое значение для любого слова из этого списка». Аналитически закон Ципфа может быть выражен в виде:

$$f \cdot r = c,$$

где f – частота встречаемости слова в тексте; r – ранг (порядковый номер) слова в списке; c – эмпирическая постоянная величина.

Полученная зависимость графически выражается гиперболой.

Позднее Б. Мандельброт предложил теоретическое обоснование закона Ципфа, основанного на эксперименте. Он полагал, что можно сравнивать письменный язык с кодированием, причем все знаки должны иметь определенную «стоимость». Исходя из требований минимальной стоимости сообщений, Б. Мандельброт математическим путем пришел к зависимости, аналогичной закону Ципфа.

Применение теории фракталов при анализе информационного пространства позволяет с общей позиции взглянуть на закономерности, составляющие основы информатики. В информационном пространстве возникают, формируются, растут и размножаются кластеры – группы взаимосвязанных сайтов. Системы, основанные на кластерном анализе, самостоятельно выявляют новые признаки объектов и распределяют объекты по новым группам. Так, компания TouchGraph разработала

оригинальный Java-апплет Google Browser, позволяющий визуализировать сложнейшие "родственные" связи между Web-сайтами. Для своей работы апплет использует механизм поиска похожих сайтов, реализованный в системе Google.

На основании закона Ципфа и, применяя Java-апплет TouchGraph Google Browser, были проанализированы такие в Web-порталы как wiki, microsoft, yandex, а также ряд Web-сайтов. Построение моделей информационного пространства и этапы анализа рассмотрены на примере Web-портала wiki, который представляет собой гипертекстовую среду (комплекс Web-сайтов) для сбора и структурирования письменных сведений.

На первом этапе была построена обобщенная модель взаимосвязей в информационном пространстве на база веб-портала wiki. Модель строится следующим образом: в поисковую систему Google отправляется запрос на получение информации о наиболее похожих сайтах, затем – о сайтах, наиболее похожих на эти сайты, и так далее. Если между сайтами на втором, третьем или последующих этапах обнаруживаются взаимосвязи, то они тоже соединяются между собой. Таким образом, создается обширная карта части сети Internet, по крайней мере, карта того участка сети, в котором находится указанный сайт. Карту части сети Internet можно представить в виде графа, где Web-страницы отображаются в виде точек, а гиперссылки – в виде линий.

На втором этапе определяется константа Ципфа. Для анализа информационного простран-ства был применен фрактальный подход, базирующийся на свойстве самоподобия информационного пространства. Свойство самоподобия выражается с помощью закона Ципфа, где f – частота встречаемости сайта; r – ранг (порядковый номер) сайта в списке всех сайтов; c – эмпирическая постоянная величина. Для Web-портала wiki были вычислены значения данных величин и определено усредненное значение константы Ципфа, примерно равное 0,26.

Аналогичные вычисления были проведены для Web-сайта www.volgograd.ru. Соответственно, константа Ципфа для данного Web-сайта получилась равной 0,257. Для Web-портала корпорации Microsoft константа имеет значение 0,256. Для поискового сайта Yandex – 0,262.

На основе проведенных исследований можно сделать вывод, что все Web-узлы информационного пространства взаимосвязаны и обладают свойством самоподобия, причем емкость информационного пространства (константа Ципфа) приблизительно равна 0,26.

На третьем этапе для доказательства данного предположения был проведен более детальный анализ одного из перечисленных Web-сайтов, а именно www.volgograd.ru. Взаимосвязи между сайтами были детализированы до более низкого уровня, также вычислялись все параметры закона Ципфа (ранг сайта, количество ссылок на сайт, частота упоминания сайта и коэффициент Ципфа для конкретного сайта). Константа Ципфа в данном случае получилась равной 0,241. Расчет константы Ципфа для расширенной модели Web-портала wiki показал результат 0,258.

Таким образом, было показано, что емкость информационного пространства лежит в диапазоне от 0,24 до 0,26 и данное утверждение справедливо для информационного пространства в целом.

Выводы

Был проведен анализ топологии информационного пространства, а также изучена сетевая навигация некоторых Web-порталов, таких как wiki, microsoft, yandex, а также Web-сайтов www.volgograd.ru, www.rambler.ru и других. Для каждого из данных Web-узлов были спроектированы модели информационного пространства на основе фрактального подхода, базирующегося на свойстве самоподобия информационного пространства. Топология и характеристики модели информационного пространства оказались схожими для различных подмножеств Web-пространства, подтверждая тем самым возможность рассмотрения Web-пространства как фрактала.