Организация и использование информационных ресурсов

ПОКРОВСКИЙ Дмитрий Николаевич *– аспирант Московского Технического Университета Связи и Информатики*

ПРИНЦИПЫ ОРГАНИЗАЦИИ ТЕМАТИЧЕСКОГО ПОИСКА В РАСПРЕДЕЛЕННОЙ ВЫЧИСЛИТЕЛЬНОЙ СРЕДЕ

Постоянный рост объемов информации в глобальной сети проявил насущную потребность в отыскании высокоэффективных механизмах поиска, обеспечивающих максимальную близость отклика системы к содержанию запроса пользователя. О продуктивности ведущейся в настоящее время активной научной и практической работы, связанной с решением этой задачи, свидетельствует бурный рост и постоянное совершенствование систем информационного поиска, составляющих новый, высокодоходный сектор рынка информационных технологий. Вместе с тем, направление для дальнейшего развития поисковых систем, таких как Google или Yahoo, в подавляющей степени связано с организацией дополнительных сервисов, способствующих покрытию все большего пространства интернета, и лишь отчасти обеспечивает поиск, лежащий в плоскости смысловой близости некоторого подмножества документов. Данное обстоятельство продиктовано позиционированием большинства современных поисковых систем, использующих механизмы полнотекстового индексирования, как ориентированных на охват всего многообразия тематик, представленного в сети информации. Вместе с тем, в условиях непрерывно ускоряющегося роста объемов сетевых информационных ресурсов проблема релевантности формируемой выборки информации становится краеугольным камнем. Особенно остро этот вопрос стоит для документов научно-технической информации. Результаты поиска должны отражать смысловую направленность исходного запроса, а не наличие у некоторой группы документов составляющих запрос терминов.

В качестве одного из возможных подходов к решению задачи тематического поиска предлагается модель распределенной вычислительной сети обмена научно-технической информацией, где за основу взята архитектура децентрализованной одноранговой распределенной сети смешанного типа [1]. При такой организации ролевое участие подключенных компьютеров делится на потребителей информации, поставщиков информации и узлы, выполняющие обслуживающие функции, в состав которых входят задачи обновления поисковых индексов, тематическая кластеризация, маршрутизация и поиск. При этом указанное деление не исключает возможности совмещения всех ролей в рамках одного узла.

Важным отличием от существующих подходов к организации поиска в распределенных сетях является замена механизма широковещательной рассылки запросов между обслуживающими узлами на использование карт маршрутизации. При этом запрос пользователя передается не от узла к узлу, а направляется непосредственно на определенный, наиболее полно ассоциированный с отвечающей исходному запросу тематикой, узел.

Для реализации подобной схемы используется механизм разбиения всей совокупности поисковых индексов на отдельные, неделимые подмножества, группирующие документы с близкой тематической направленностью, которые затем распределяются по обслуживающим узлам сети. На рисунке 1 изображена обобщенная структура процесса поиска в сети научного знания.

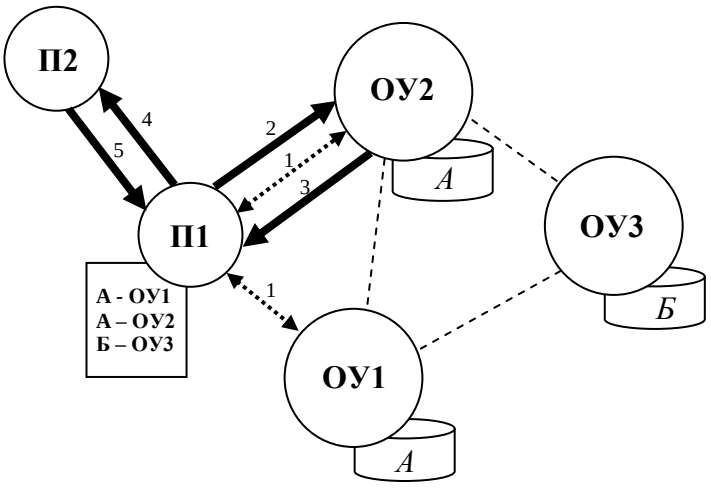
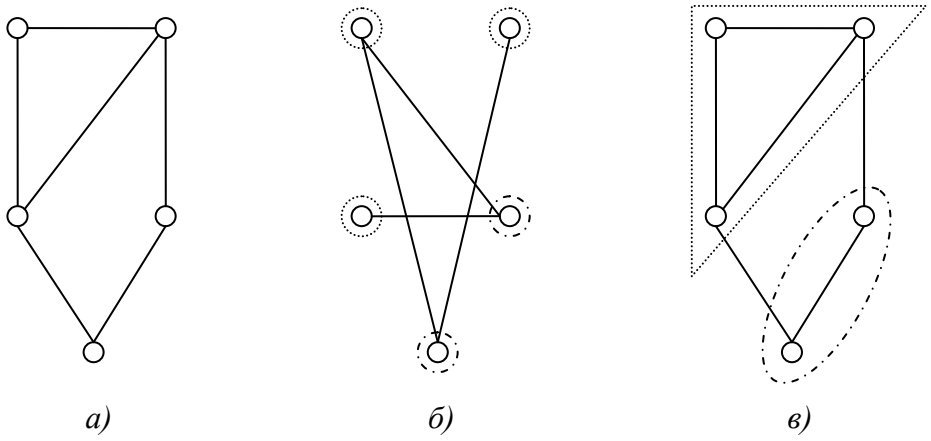


Рис. 1. Обобщенная структура процесса поиска в сети научного знания.

Пользователь П1 (потребитель информации) формирует поисковый запрос по тематике «А», который может быть передан ассоциированным с данной темой обслуживающим узлам ОУ1 и ОУ2. Основываясь на контрольных данных (1) возвращенных узлами (время отклика, загруженность системы и т.д.), выбирается ОУ2 (2). Затем производится поиск в сегменте базы индексов документов, отвечающих указанной теме, результаты ранжируются, и сформированный узлом ОУ2 список ссылок возвращается пользователю (3). В приведенном примере искомая информация размещена у пользователя П2 (поставщика информации), и пользователь П1 инициирует прямое соединение (4) с ним для получения требуемого документа (5).

В отличие от классической модели библиотечно-информационной системы, где реализуется статическая иерархия рубрик, в сети научного знания используется процесс динамического связывания, основанный на смысловых признаках, идентифицирующих информационный источник [2]. В целях однозначной интерпретации семантической принадлежности предполагается использование вербальной системы Государственного рубрикатора научно-технической информации. Каждому из M документов множества $D = \{d_1, ...d_M\}$ сопоставляется N - мерный вектор, получаемый в результате автоматической рубрикации по N рубрикам [3] и отражающий степень принадлежности документа каждой из них. Затем в сформированном массиве определяется близость каждой пары документов и на основании этих данных строится неориентированный граф G, в котором вершинами выступают документы сети, а ребрам соответствуют существенные (большие некоторого заданного порогового значения) взвешенные связи между документами, как это представлено на рисунке 2а.



а) *б)* *в)*

Рис. 2

Факторизация ребер графа по варьируемому пороговому значению позволяет исключить несущественные, малозначимые связи между документами. Полученный граф становится исходной моделью описания смыслового пространства множества документов. Для выявления устойчивых семантических связей предлагается использование приближенного алгоритма решения хроматической задачи по отношению к дополнению графа G, представляющему (на рисунке 2б) структуру незначимых и малозначимых связей. Приближенный алгоритм основан на выделении независимых подмножеств вершин (которые «окрашиваются в один цвет») с помощью «жадного» определения наименьшего вершинного покрытия [4] .

Таким образом, определяются подструктуры исходной совокупности, обладающие 1-достижимостью по существенным связям и отражающие близкую смысловую направленность выделенного подмножества документов (рис. 2в). Это позволяет вести поиск не во всем пространстве представленных тематик, а лишь в определенной его части, обусловленной семантическими связями, что, совместно с использованием поискового образа, призвано обеспечить высокий уровень релевантности результатов возвращаемых системой.

Результатом работы описанной процедуры будут являться тематические домены (подмножества поисковых индексов), которые размещаются на обслуживающих узлах по следующему правилу: определяется минимальное необходимое число узлов, на которые могут быть равномерно загружены все сформированные домены, затем полученный набор последовательно реплицируются на остальных обслуживающих узлах сети (рис. 3).

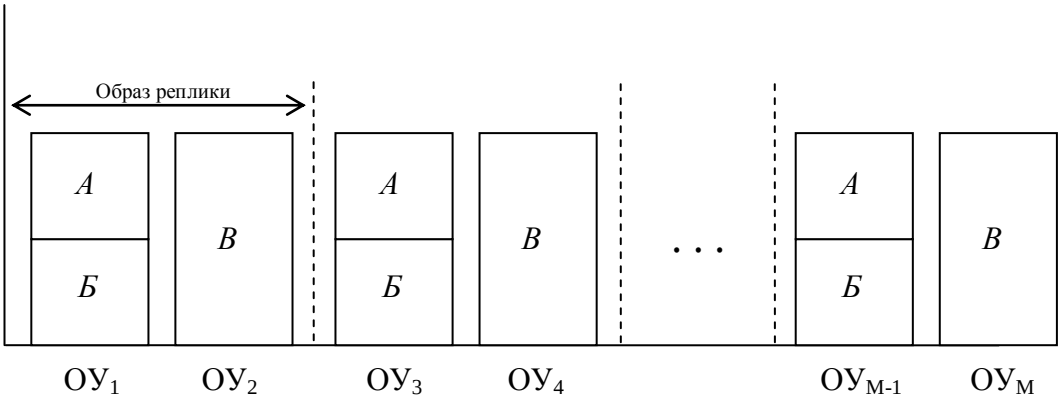


Рис. 3

Введение избыточности при такой схеме организации позволяет управлять соотношением уровня отказоустойчивости системы (за счет увеличения числа дублирующих обслуживающих узлов) и производительности (за счет увеличения уникальных обслуживающих узлов, уменьшения объема базы поисковых индексов в пересчете на один узел).

В целях отслеживания динамики изменений в структуре семантических взаимосвязей между документами для проведения своевременной реконфигурации обслуживающих узлов и маршрутных карт процедура разбиения должна периодически повторяться.

Применение предлагаемой организации поисковых механизмов в сочетании с современными методами лексического анализа документов [3] отличает также возможность повышения уровня релевантности возвращаемых данных в процессе отыскания нужного пользователю документа. Указанная возможность связана с тем, что предварительная настройка системы на стороне клиента может включать в себя лексический анализ ранее найденных документов, отражающих более полно (чем запрос) тематическую направленность предполагаемого поиска, что дает возможность пользователю точнее выйти на обслуживающий узел. Данное действие, несмотря на ухудшение эргономических характеристик пользовательского интерфейса, обеспечивает возможность автоматизировать локализацию соответствующего тематического домена, в рамках которого будет вестись поиск.

Широкое распространение одноранговых децентрализованных сетей, включая научную и образовательную сферу, объясняется эффективностью данного подхода к формированию коммуникационной среды, обладающей свойством самоорганизации, функционирующей поверх существующей сетевой инфраструктуры, позволяющей вовлекать ее участников в активный процесс информационного обмена. В свою очередь, реализация механизма поиска документов по метаданным, описывающим принадлежность к определенной области знаний или тематическому домену, должна значительно раздвинуть рамки возможностей организации научно-исследовательских сообществ, ведущих работы по какой-либо проблематике, предлагая удобный и эффективный инструмент обмена знаниями.

Литература

1. Steve Waterhouse, David M. Doolin, Gene Kan, Yaroslav Faybishenko Distributed Search in Peer-to-Peer Networks // IEEE Internet Computing, Jan-Feb 2002.

2. Малахов А. А., Мелихов В. О., Качак В. В., Комарова Е. А. Принципы формирования научных приоритетов высшей школы с использованием аналитического и экспертного прогнозирования // Научно-техническая информация. Сер. 1. - 2002 – №11. - С. 20-24.

3. Н.С. Соловьева, Н.В. Сомин. Опыт реализации лексикостатистического метода рубрицирования текстовых сообщений // Системы и средства информатики. - 2000. - Вып. 10. - С.205-215.

Пападимитриу Х., Стайглиц К. Комбинаторная оптимизация: алгоритмы и сложность. – М.: Мир, 1985. - С. 418-421.