УДК 004.3

И.А. Коломойцева, ст. преподаватель,
Н.Н. Дацун, канд. ф.-м. наук, доц.
Донецкий национальный технический университет
kolomoit@r5.dgtu.donetsk.ua

Алгоритм работы брокера ГРИД-системы для решения задачи метапоиска

В данной работе рассматривается решение задачи метапоиска с помощью ГРИД-системы. Разработана общая схема ГРИД-системы для решения этой задачи. Предложен алгоритм работы брокера ГРИД-системы, который по набору запросов получает множество выдач стандартных поисковых машин. Использование такого подхода позволит обойти существующее ограничение поисковых машин.

Ключевые слова: ГРИД-система , брокер, метапоиск.

Введение

Интернет стремительно развивается. По данным сайта www.netcraft.com количество сайтов в 2004 году превышало 50 млн., в 2011 году - 270 млн., в феврале 2013 года - 630 млн. Если темпы увеличения сайтов в 2004 году составляли примерно 1 млн. сайтов в месяц, то в феврале 2012 года составили 8,2 млн. Увеличивается также количество документов в сети [1].

Такое количество документов делает важной задачей задачу поиска информации в сети.

В последнее время для решения этой проблемы активно разрабатываются метапоисковые системы [2].

Метапоисковая система (метапоисковая машина) — это поисковая система, которая в отличие от классических поисковых машин не имеет собственной базы данных и собственного поискового индекса, а формирует поисковую выдачу за счет смешивания и переранжирования результатов поиска других поисковых систем [2].

Среди разработанных метапоисковых систем следует отметить Web Scout (ищет новости, конференции, аукционы), 1 SECOND, search.da.ru, exactus [2].

Задача метапоиска ресурсозатратная, обладает естественным распараллеливанием. Кроме этого, при реализации автоматических запросов к стандартным поисковым машинам возникают проблемы, связанные с ограничением, установленным этими машинами. Они ограничивают количество автоматических запросов, которые можно выполнить в единицу времени с одного IP-адреса.

Снять это ограничение и увеличить скорость работы метапоисковой системы можно помощью ГРИД-системы.

Направление развития ГРИД-технологии происходит по трём направлениям [3]:

1) вычислительные ГРИД;

2) семантические ГРИД;

3) ГРИД для интенсивной обработки данных.

ГРИД первого направления предназначены для решения вычислительных задач, например с использованием численных методов (решение уравнение математической физики) [3].

Семантические ГРИД предназначены для оперирования данными из различных баз данных.

ГРИД третьего направления предназначены для обработки огромных объёмов данных несложными программами, которые можно реализовать на персональном компьютере. Сложность этого направления – доставка данных для обработки и пересылка результатов [3].

ГРИД-системы используются для решения задач прогноза погоды, обработка космических, астрофизических, геофизических данных и для БАК, индексирование новостных БД, в физике высоких энергий, биомедицинских науках, вычислительной химии, ядерного синтеза, финансово-экономических исследований [3, 4, 5, 6, 7].

Задача метапоиска требует обработки большого объёма информации. Поэтому её решение с помощью ГРИД-системы является актуальной задачей.

Общая постановка задачи

ГРИД-система для решения задачи метапоиска делится на следующие подсистемы (рисунок 1):

– подсистема интерфейса пользователя, которая обеспечивает ввод запроса от пользователя и вывод результатов поиска;

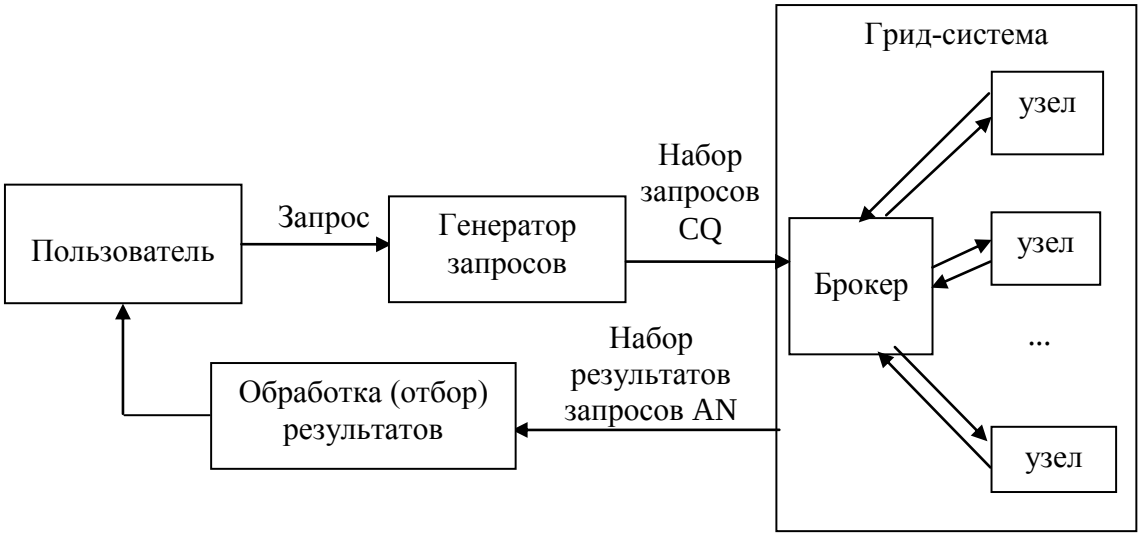

Рисунок 1 – Общая схема ГРИД-системы, решающей задачу метапоиска

– подсистема генерации запросов, которая по введенному пользователем генерирует запросы, семантически идентичные исходному;

– подсистема работы брокера (планировщика), которая распределяет запросы между узлами и принимает результат работы;

– подсистема работы узлов Грид-системы, выполняющих поиск;

– подсистема отбора релевантных результатов, которая получает их от брокера и после обработки передаёт подсистеме интерфейса пользователя.

Базой для генерации запросов для подсистемы генерации запросов является тезаурус. В работе [8] предложена объектная модель семантического анализа естественно-языкового текста, используемая для тезауруса запросов в таких предметных областях, как "заболевания" [8] и "лекарства" [9].

Узлы ГРИД-системы при решении задачи метапоиска могут выполнять две задачи:

- запускать стандартную поисковую машину с конкретными запросами и получать результат;

- в БД текстов, которая хранится на узлах, осуществлять поиск по конкретному запросу и выдавать список документов, ему удовлетворяющих.

В [10, 11] предлагается структура подсистемы, которая на основе объектной модели из [5], выполняет анализ текстовой БД и отбирает нужные документы. Отобранные документы вместе с результатами работы стандартных поисковых машин, отправляются в подсистему подсистема отбора релевантных результатов. Работа брокерами с узлами, работающими с полнотекстовой БД, не является предметом данной статьи.

Отбор результатов в подсистеме подсистема отбора релевантных результатов может осуществляться на основе двух подходов:

– объектного;

– функционального.

В работах [8-11] описан метод отбора релевантных результатов на основе анализа выдач с помощью тезауруса, построенного посредством объектной модели.

Альтернативой является функциональный подход, представленный в работе [11]. В соответствии с этим подходом выдачи результатов запросов можно представить в виде функций, аргументы которых должны быть сопоставлены с образцами. Однако этот подход имеет более высокую вычислительную сложность.

Описание алгоритма работы брокера

Алгоритм работы брокера включает две процедуры:

1) **PFQ** – формирование множества узлов, которые будут решать задачу метапоиска, и множества запросов, отправляемых на конкретный узел;

2) **PFA** – формирование множества результатов выполнения запросов.

Описание PFQ .

Входные данные.

Входными данными являются:

1) множество узлов ГРИД-системы;

2) множество запросов, поступающих от подсистемы генерации запросов.

Множество узлов, которые входят в ГРИД-систему обозначим PN. Количество этих узлов – n1.

Множество узлов PN можно описать набором из элементов:

$$PN =< V_i, Fr_i, S_i, CN_i, R_i >, i = \overline{1, n1},$$

где

Fr_i – тактовая частота процессора на узле i;

V_i – объём оперативной памяти (ОП) на узле i;

S_i – объём дискового пространства на узлеi;

CN_i – пропускная способность сетевой карты узла i;

R_i – ранг узла i.

Ранжирование узлов выполняется по следующему принципу:

1) по частоте ЦП узлов;

2) по объёму ОП;

3) по объёму дискового пространства.

Ранг присваивается узлу в момент регистрации в ГРИД-системе.

Множество запросов, поступающих брокеру ГРИД-системы от подсистемы генерации заявок, обозначим CQ, а количество запросов, поступивших от подсистемы генерации запросов – n2. CQ можно описать набором элементов:

$$CQ =< i, Tq_i >, i = \overline{1, n2},$$

где

Tq_i – текст запроса.

Выходные данные.

Выходными данными является множество узлов Res, на которых будет решаться задача метапоиска, с заданием конкретных параметров поиска.

Порядок действий PFQ.

1 шаг. Определение множества ND – узлов, которые могут участвовать в решении задачи метапоиска. В ND попадают те узлы из PN, дисковое пространство которых позволит сохранить результаты поиска. Обозначим узлы, не подходящие для решения задач по размеру дискового пространства, как NNN. Тогда $ND = PN - NNN.$ Количество узлов в ND обозначим как n3: $n3 \le n1.$ ND имеет такую же структуру, как PN:

$$ND =< V_i, Fr_i, S_i, CN_i, R_i >, i = \overline{1, n3}.$$

Обозначим функцию, которая выполняет отображение множества PN на множество ND, как F1. Тогда ND можно определить как:

$$ND = \{nd \mid F1 : pn_i \to nd_j\}, i = \overline{1, n1}, j = \overline{1, n1}.$$

F1 определяется следующим образом:

$$F1 : \exists p : (nd_p = pn_i,$$

$$pn_i.s \ge n2 \cdot mz), i = \overline{1, n1}, p = \overline{1, n1},$$

где

mz – константа, равная количеству байтов, необходимому для хранения результата выполнения одного запроса; $mz=10^7$ Б.

2 шаг. Формирование множества узлов Res, на которых непосредственно будут выполняться запросы:

$$Res =< V_i, Fr_i, S_i, CN_i, R_i, Q_i, T_i >, i = \overline{1, n4},$$

где

$$n4 = \begin{cases} n2, \text{если } n2 \le n3 \\ n3, \text{иначе} \end{cases}$$

где

Q_i – количество запросов, выполняющихся на узле i;

Ti – тексты запросов, выполняющихся на узле i.

Res формируется с помощью функции F2. Для её задания нужно определить количество узлов n5, которые решают конкретную задачу метапоиска. n5 определяется по формуле:

$$n5 = \begin{cases} 1, \text{если } n2 \le n3 \\ f_{div}(n2, n3) + f_{\dot{-}sg}(f_{mod}(f_{\dot{-}}(res_j.r, \\ f_{div}(n2, n3))), \text{иначе} \end{cases}$$

f_{div} – целая часть от деления.

f_{mod} – остаток от деления.

$f_{\dot{-}}$ – арифметическая разность. Определяется следующим образом:

$$f_{\dot{-}}(x, y) = \begin{cases} x - y, \text{если } x > y \\ 0, \text{иначе} \end{cases}$$

$f_{\dot{-}sg}$ – антизнак. Определяется следующим образом:

$$f_{\dot{-}sg}(x) = \begin{cases} 0, \text{если } x > 0 \\ 1, x = 0 \end{cases}$$

Тогда функция F2 определяется следующим образом:

$$F2 : \exists j : (res_j.v = nd_i.v,\ res_j.fr = nd_i.fr,$$

$$res_j.s = nd_i.s,\ res_j.cn = nd_i.cn,\ res_j.r = nd_i.r,$$

$$res_j.q = n5,\ res_j.t = \{cq_k.tq \mid k = \overline{1, n5}\}), i = \overline{1, n4},$$

$$j = \overline{1, n4}.$$

После определения нагрузки для j-узла из множества CQ удаляются запросы, отправленные на этот узел.

Описание PFA .

Входные данные.

Входными данными являются результаты работы узлов.

Данные, которые возвращает i-тый узел брокеру ГРИД-системы, обозначим AN_i:

$$AN_i = < Tq_j, Re\, q_j >, j = \overline{1, n6}, i = \overline{1, n4},$$

где

Req_i – это набор результатов работы стандартных поисковых машин, который содержит следующую информацию:

- заголовок выдачи;
- краткое описание найденного результата;
- ссылку.

n6 – количество выдач, пришедших от i-того узла, $n6=n5 \cdot 100$.

Выходные данные.

Множество наборов результатов AN, полученных от стандартных поисковых машин.

Порядок действий PFA.

Общее множество результатов (выдач) формируется путём объединения результатов, пришедших с i-тых узлов.

$$AN = \bigcup_{i=1}^{n4} AN_i .$$

Сложность алгоритма зависит от n2. В свою очередь n2 зависит от структуры исходного запроса (количества слов, наличия устойчивых словосочетаний и т.п.).

Заключение

В данной работе рассмотрена организация ГРИД-системы, решающей задачу метапоиска. В частности выделены основные её структурные элементы.

Описан алгоритм работы брокера ГРИД-системы, включающий процедуры формирования задачи узлам и формирования ответа Грид-системы.

Дальнейшее развитие предложенного алгоритма состоит в повышении его быстродействия путём уменьшения времени получения результатов за счёт учёта географического расположения узлов ГРИД-системы.

Список литературы

1. Ландэ Д.В. Поиск знаний в Internet. Профессиональная работа.: пер. с англ. / Д.В. Ландэ. – М.: Издательский дом "Вильямс", 2005. – 272 с.
2. Пасько В.П. Энциклопедия ПК. Аппаратура. Программы. Интернет / В.П. Пасько. – Киев, издательская группа BHV; СПб: Питер, 2004. – 800 с.
3. Котляр В.В. Применение Грид-технологий для задач интненсивной обработки данных / В.В. Котляр // Электронный доступ: http:litcey.ru/geografiya/5431/index.html
4. Куссуль Н.Н. Grid-технологии в системах мониторинга окружающей среды / Е.А. Лупян, А.Ю. Шелестов, Л. Глухи, П. Копп // Современные проблемы дистанционного зондирования Земли из космоса, сборник научных статей. – 2008. – Выпуск 5, том 2-й. – С. 538-547.
5. Castellano M. Biomedical Text Mining Using a Grid Computing Approach / M. Castellano, G. Mastronardi, Decataldo G., Pisciotta L., Tarricone G., Cariello L., Bevilacqua V. // Электронный доступ: http://cdn.intechopen.com/pdfs-wm/12928.pdf.
6. Hughes B. Grid-based Indexing of a Newswire Corpus / B. Hughes, S. Venugopal, R . Buyya // Электронный доступ: http://www.cloudbus.org/papers/nlp-newswire-grid.pdf.
7. Li Q. The Future-Oriented Grid-Smart Grid / Q. Li, M. Zhou // Journal of computers, Vol 6. – №1. – 2011. – p.98-105.
8. Коломойцева И.А. Объектная модель семантического анализа естественно-языкового медицинского текста / И.А. Коломойцева // Научные труды Донецкого национального технического университета. Серия «Информатика, кибернетика и вычислительная техника». – 2007. – Выпуск 8 (120). – С. 141-150.
9. Коломойцева И.А. Компьютерный тезаурус для предметной области «ЛЕКАРСТВА» / И.А. Коломойцева // Моделирование и компьютерная графика: материалы 4-й международной научно-технической конференции, г. Донецк, 5-8 октября 2011 г. – Донецк: ДонНТУ, Министерство науки и образования, молодёжи и спорта, 2011. – С. 161-165.
10. Коломойцева И.А. Объектная модель естественно-языкового медицинского текста на примере системы «ФармАналитик» / И.А. Коломойцева // Научные труды Донецкого национального технического университета. Серия «Информатика, кибернетика и вычислительная техника». – 2009. – Выпуск 10 (153). – С. 303-306.
11. Коломойцева И.А. Функциональная модель медицинского естественно-языкового текста / И.А. Коломойцева // Научные труды Донецкого национального технического университета. Серия «Информатика, кибернетика и вычислительная техника». – 2008. – Выпуск 9 (132). – С. 237-241.

Надійшла до редакції 10.04.2014

І.О. КОЛОМОЙЦЕВА, Н.М. ДАЦУН
Донецький національний технічний університет

АЛГОРИТМ РОБОТИ БРОКЕРА ГРІД-СИСТЕМИ ДЛЯ ВИРІШЕННЯ ЗАДАЧІ МЕТАПОШУКУ

В даній роботі розглядається рішення задачі метапошуку за допомогою ГРІД-системи. Розроблено загальну схему ГРІД-системи для вирішення цього завдання. Запропоновано алгоритм роботи брокера ГРІД-системи, який за набором запитів отримує безліч видач стандартних пошукових машин. Використання такого підходу дозволить обійти існуюче обмеження пошукових машин.

Ключові слова: ГРІД-система, брокер, метапошук.

I.A. KOLOMOITSEVA, N.N. DATSUN
Donetsk National Technical University

ALGORITHM OF THE GRID-SYSTEM BROKER FOR SOLVING METASEARCH TASKS

The paper considers the solving of the problem of metasearch by using GRID. The General scheme of the GRID to solve this problem was developed. This scheme includes the subsystem of user interface, the subsystem of generating of requests, the subsystem of the broker, the subsystem of nodes, the subsystem of selection of relevant results. The subsystem of user interface is designed for user interaction with the grid system. The subsystem of request generation creates a lot of queries by the user's request. The subsystem of the broker distributes requests between the nodes and takes the result of the work. The subsystem of nodes executes a search using the standard search engines. The subsystem selection of relevant results selects relevant results from a variety of responses that came from the broker. In this article the algorithm of the broker of GRID system is offered. It includes two procedures. The first procedure performs two actions. First, it generates a list of nodes that will solve a particular task search. Second, it generates a list of queries that will run on those nodes. The second procedure collects the results the work of the nodes and generates a common set of results. Such an approach will bypass the existing limitation of search engines. This limitation determines the maximum number of requests per unit time.

Keywords: GRID system, broker, metasearch.