Специализирани търсещи машини в Интернет пространството – основни принципи и архитектура

Ирена Вълова, Йордан Калмуков

Abstract: *This paper describes the main principles and problems of building an automated tool for searching the web. Choosing the most appropriate search strategy seems to be the key point for achieving higher efficiency, denoted as the number of useful downloaded resources per time unit.*

Converting the directed weightless graph of WWW to a directed weighted one is absolutely necessary to find an adequate way for navigating through the web. The accuracy of weights' calculations directly determines how efficient the web crawler will be. Several ways of calculating the weight of vertices and edges are suggested and briefly described in this paper. The paper also proposes a common architecture of web spider designated to find and store images.

Key words: *web crawler, web spider, offline browser, automated tool for searching images on the web, searching strategy.*

ВЪВЕДЕНИЕ

Развитието на информационните технологии и съвременните компютърни системи както софтуерно, така и хардуерно дават възможност за тяхното приложение във всички области, в които информацията позволява цифрово представяне, а също така, позволява съхранение и организиране на огромни количества различна по тип и структура информация, което води до развитието на базите от данни и информационните системи от една страна и до стремеж към усъвършенстване на услугите, предлагани при обслужване на потребителите на тази информация, от друга. Благодарение на това развитие в последните години са изключително актуални бази от данни, работещи не само с текстова информация, но и съхраняващи графични, аудио и видео данни. Съвсем естествено възниква проблемът за събирането, организацията и структурирането на цялата тази информация, за да може тя да бъде полезна за всички потребители и съответно при използване на новите типове данни, при описанието и структурирането на информацията, да се избегне максимално влиянието на субективния фактор. Тоест стремеж към автоматизиране на процеса на организация и достъп до тази информация.

Решение на този проблем при съвременното количество информация в Интернет и нейното ежедневно обновяване и коригиране е използването на търсещи роботи или паяци за съответен тип данни.

ОСНОВНИ ТЕОРЕТИЧНИ ПОСТАНОВКИ

Търсещите роботи в Web са ***програми, които използват графовата структура на Web пространството, за да го обхождат страница по страница.*** Този тип програми са известни още като паяци, пътешественици, червеи и други подобни понятия, които са дошли преди всичко от същността на изпълняваните от програмата функции.

Web роботите възникват с цел да се осигури организация и търсене на web страници на база на някакви техни представяния, съхранени в локални бази от данни. В последствие тези бази от данни могат да бъдат използвани за нуждите на различни приложения.

Ако web пространството представляваше статично множество от сайтове, то това би опростило значително функциите и задачите на тези търсещи роботи, защото веднъж обходили всички страници и създали базата от данни със съответната информация, няма да бъде необходимо отново да обхождат и претърсват многократно. Web пространството обаче, е твърде динамично, което налага усложнение на функциите на web роботите, с цел поддържане на цялостност и актуалност на информацията в съответните бази от данни [1]. Стремежът да се

реализира такъв тип програма, поддържаща абсолютна изчерпателност на резултатите във всички области, води до тромаво работещо приложение. За да се избегне това, се използват web роботи с определена селективност по отношение на страниците, които обхождат или по отношение на търсената от тях информация. Това са така наречените евристични роботи, използващи тематично или фокусирано обхождане и търсене [2, 3, 4, 5]. Използват се най-вече като комбинация с търсещи машини за различен тип информация.

Основният проблем при автоматизираното претърсване на Интернет е създаването на адекватна стратегия за обхождане на страниците, гарантираща достатъчно висока ефективност на търсенето. Скоростта на публикуване на нови ресурси в Интернет е толкова голяма, че постигането на резултати с висока качествена и количествена стойност за минимално време е от решаващо значение за всеки паяк.

АЛГОРИТЪМ НА ДЕЙСТВИЕ И ОБОБЩЕНА АРХИТЕКТУРА

По своята същност ***услугата WWW представлява цикличен ориентиран безтегловен граф с краен, но все пак безумно голям брой възли.*** Сляпото обхождане на подобна структура не само е безсмислено, но поради огромния брой възли е и невъзможно в рамките на разумен времеви интервал. Ето защо обхождането на интернет страниците трябва да бъде насочвано по някакъв начин, за да се постигне спомената вече висока ефективност. С други думи необходимо е да се въведат критерии за оценка на перспективността на избраната посока на обхождане. Числеността изражение на тази оценка следва да се използва като тегло на възлите и/или дъгите. По този начин WWW от безтегловен се преобразува в претеглен граф. Това преобразуване е абсолютно необходимо за да се реализира т.нар. динамично само насочване на търсенето, т.е. алгоритъмът, изработващ стратегията за търсене да може във всеки един момент сам да вземе решение с кой възел в графа да продължи. Като алгоритъм за изработване на стратегията за обхождане успешно може да се приложи някой от добре познатите от изкуствения интелект евристични алгоритми за осведомено търсене в граф. По-голям интерес при проектирането на паяка представлява начинът, по който ще се изчисляват тегловните коефициенти, тъй като това именно са данните, които пряко влияят върху вземането на решение и като такива тяхната адекватност е от съществено значение за постигане на заветната цел – точност и ефективност.

Най-общо начините за изчисляване на теглата на възлите са следните:

- паякът сам по определени критерии изчислява теглата;
- теглото на даден възел се представя като сума от теглата на всички сочещи към него дъги;
- комбинация от горните два начина.

В ранните години на Интернет, търсещите машини изчисляваха ранга на страниците предимно по втория начин – като сума от теглата на дъгите сочещи към дадения възел (в случая страница). И тъй като по него време за изчисляване на тегла на дъгите малцина са мислили, то колко по-напред ще бъде дадена страница в резултатите от търсенето зависеше преди всичко от това колко хипер връзки сочат към нея. Знаейки това много предприемчиви търговци създаваха мрежи от по 20-50 абсолютно еднакви електронни магазина, които се различаваха само по имената и графичния дизайн, но от всеки един от тях имаше връзки към всички останали и това даваше добър резултат за визуализиране на дадения сайт по-нагоре в списъка на резултатите от търсенето.

С течение на времето този начин за изчисляване на ранга на страниците започна да избледнява и да отстъпва място на други по-адекватни начини, базиращи се на идеята, че паякът може сам по време на работа да оценява важността на всеки сайт по определени критерии – примерно не само колко сайта

сочат към него, но и каква е важността на сочещите към него, каква е тяхната честота на актуализация на информацията, каква е тяхната посещаемост и т.н. И не на последно място ако става дума за тематично насочено търсене доколко даденият сайт / web ресурс отговаря на тематиката и други предварително дефинирани изисквания.

Теглата на дъгите отразяват степента на някаква семантична близост между възлите. Тогава стойността на теглото на дадена дъга ще зависи от тематичната близост на възлите, които свързва. Оценката на тематичната близост на два възела не е предмет на този доклад, но е редно да се спомене че в повечето случаи тя се намира чрез анализ на семантиката на текстовото съдържание на страницата, а ключовите думи дефинирани в <HEAD> частта на HTML документа се използват рядко (поради възможността за умишлено заблуждаване на търсещите машини). Възможно е теглото на дадена дъга да зависи не само от тематичната близост на възлите, които свързва, но също и от техните тегла.

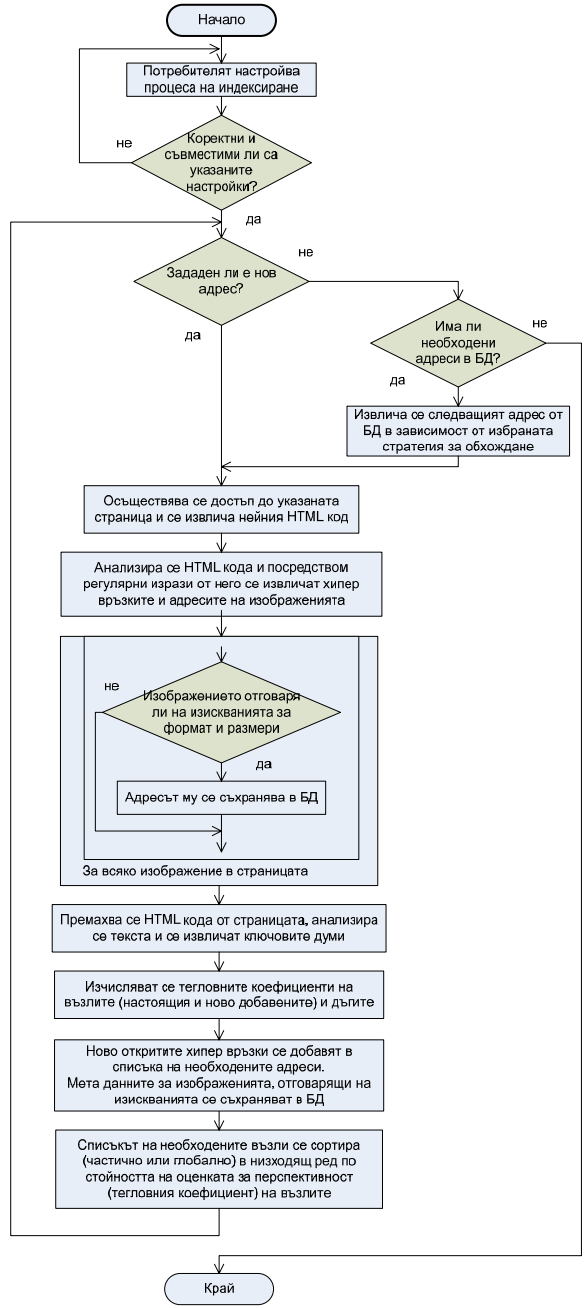
Всеки паяк поддържа своя собствена структура от данни, съдържаща адресите на не обходените все още възли (или на вече обходени, но подлежащи на ре индексиране). За простота тази структура ще бъде наричана списък, но практически може да е и по-сложна структура. Целта на алгоритъма за изработване на стратегията за обхождане е преди всичко да изчисли по най-адекватния възможен начин стойността на тегловните коефициенти на възлите и дъгите. Останалото се състои в просто сортиране на списъка в низходящ ред по значението на избраната оценка за перспективност на възлите. Тъй като списъкът с адреси е динамична структура, то алгоритъмът за изработване на стратегията трябва да се изпълнява за всеки ново добавен елемент към него.

На фигура 1 е представен обобщен алгоритъм на работата на паяк, предназначен за автоматично извличане и съхраняване на адресите на графични изображения, отговарящи на определени изисквания за формат и размери.

В началото списъкът с адреси е празен и потребителят трябва да зададе стартов адрес. Осъществява се достъп до него и се извлича съдържанието на страницата заедно с HTML кода. Откриват се всички хипер връзки в него и се съхраняват временно в оперативната памет. По същия начин се откриват и адресите на изображенията в страницата.

Осъществява се достъп до изображенията и се анализират техните формат и размери. В случай, че отговарят на изискванията, адресите им се съхраняват временно в оперативната памет докато се извлекат и мета данните за тях. След това от съдържанието на страницата се премахва HTML кода, така че да остане само текста, предназначен за потребителя. Този текст се анализира и от него се извличат ключовите думи описващи и страницата и изображенията в нея. Дадена дума може да се приеме за ключова в зависимост от семантиката й, граматичното й значение, честотата на срещане в текста и указаната от потребителя нейна значимост. Точно ключовите думи представляват мета данните за изображенията.

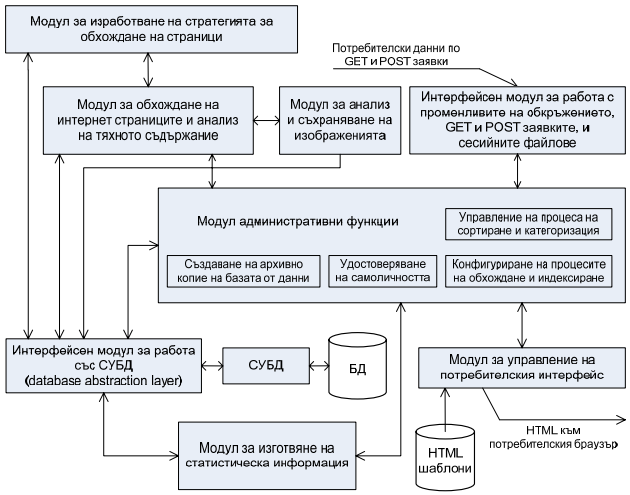
След като и те са извлечени то те заедно с адресите на изображенията, отговарящи на условията за формат и размери, се съхраняват в базата от данни (БД) на паяка. Извлечените ключови думи, както и много други фактори, също могат да участват при формирането на оценката за значимост на възела. Затова точно тук, след анализа на текста, се изчисляват тегловните коефициенти на възлите в графа. След което те заедно с адресите на откритите хипер връзки се добавят в списъка на не обходените възли. В последствие списъкът се сортира (частично или пълно) в низходящ ред по значението на тегловните коефициенти и се продължава с първия адрес, намиращ се в началото на сортирания вече списък. Анализирането на текста и извличането на мета данни за изображенията са доста времеемки операции, но тяхното извършване е абсолютно наложително, за да могат автоматично да се опишат и категоризират извлечените изображения.



Фигура 1. Обобщен алгоритъм на работа на паяк, предназначен за автоматично извличане на адресите на графични изображения, отговарящи на определени изисквания за формат и размери.

В противен случай информационният хаос ще е пълен. Т.е. паякът ще е смъкнал огромно количество изображения, но какво от това след като няма начин автоматизирано да се разбере съдържанието и семантиката на всяко едно конкретно изображение и да се направи категоризация с цел организация на изображенията на база метаданни за тях.

На фигура 2 е представена архитектурата на web-базирана система за автоматично претърсване на Интернет страници и извличане на изображения, отговарящи на определени изисквания за формат и размери.



Фигура 2. Архитектура на web-базирана система за автоматично извличане на графични изображения от Интернет.

Логиката на паяка се реализира от:

- ***Модул за обхождане на интернет страниците и анализ на тяхното съдържание***;
- ***Модул за изработване на стратегията за обхождане на страници***;
- ***Модул за анализ и съхраняване на изображенията***;
- ***Модул за изготвяне на статистическа информация***;
- И подмодулите за Управление на процеса на сортиране и категоризация и Конфигуриране на процесите на обхождане и индексиране от ***модула за Административни функции***.

Модулът за управление на потребителския интерфейс позволява напълно да се раздели бизнес логиката на приложението от потребителския му интерфейс. Това повишава гъвкавостта и дава възможност в последствие лесно и бързо да се модифицира или изцяло промени графичният интерфейс на паяка. Последният е организиран в йерархично структурирани HTML шаблони.

Модулът за работа със СУБД предоставя още едно ниво на абстракция за работа с базата от данни. Заявките подавани към него могат да бъдат на SQL; под формата на параметри на функция; или в комбинация от двете. Една от основните му цели е да преобразува стандартния SQL синтаксис на подаваните към него заявки до специфичния SQL за конкретното СУБД. Модулът за работа със СУБД предоставя унифициран начин за извличане и работа с данните независимо от СУБД-то и специфичния SQL синтаксис, който поддържа.

Друга основна цел на модула е да предостави удобен и гъвкав механизъм за по-лесно и автоматизирано обработване на извлечените от БД данни. Примерно, ако се укаже броя записи на страница, които трябва да се визуализират, модулът

автоматично изчислява броя на страниците и параметрите за навигация между тях (т.е. параметрите за връзките: първа страница | предишна страница | следваща страница | последна страница).

След като потребителят е настроил паяка, процесите на обхождане и индексиране, и е задал стартов адрес модулът за обхождане на интернет страниците взема съответните данни от БД и започва своята работа – осъществява достъп до всяка страница, анализира нейния HTML код и извлича от него хипер връзките, адресите на изображенията и ключовите думи в страницата. Адресите на изображенията се предават на модула за анализ и съхраняване на изображенията, който проверява техните формат и размери, и решава дали отговарят на изискванията и трябва да бъдат съхранени в локалната база от данни. Хипер връзките заедно с ключовите думи се предават на модула за изработване на стратегията за обхождане, който изчислява теглата на текущия и на ново откритите възли (интернет страници) в графа, добавя новите възли в списъка на необходените и го сортира (частично или пълно, в зависимост от потребителските настройки и избраната стратегия за обхождане) в низходящ ред по значението а изчисления вече коефициент на перспективност на възлите.

След обработването на текущия възел, модулът за обхождане продължава със следващия елемент от сортирания вече списък на необходените възли и т.н.

ЗАКЛЮЧЕНИЕ

Огромното количество информация и динамичната същност на интернет пространството налагат необходимостта от постоянна поддръжка и обновяване на информацията в web базираните системи за търсене и достъп до информация.

Реализацията на предложената архитектура на web робот за събиране на изображения би била изключително полезна при автоматичната организация на информацията в базите от данни от изображения. Определяща във всички случаи си остава стратегията на търсене – от нея зависи скоростта и ефективността на работа на web робота. Тъй като web пространството може да се моделира с помощта на графи, то приложими са различни модификации на алгоритмите за обхождане и търсене в претеглен граф. Като евентуална бъдеща насока на развитие на тези приложения може да се мисли за тяхната оптимизация чрез разширяването им с модул за оценка на достоверността на намерените резултати.

ЛИТЕРАТУРА

[1] A. Arasu, J. Cho, H. Garcia-Molina, A. Paepcke, and S. Raghavan. Searching the Web. ACM Transactions on Internet Technology, 1(1), 2001.

[2] S. Chakrabarti. Mining the Web. Morgan Kaufmann, 2003.

[3] J. Cho, H. Garcia-Molina, and L. Page. Eficient crawling through URL ordering. Computer Networks, 30(1-7):161-172, 1998.

[4] S. Chakrabarti, M. van den Berg, B. Dom. Focused crawling: A new approach to topic-specific Web resource discovery. Computer Networks, 31(11-16):1623-1640, 1999.

[5] A.K. McCallum, K. Nigam, J. Rennie, K. Seymore. Automating the construction of internet portals with machine learning. Information Retrieval, 3(2):127-163, 2000.

За контакти:

д-р Ирена Вълова, Катедра “Компютърни системи и технологии”, Русенски университет “Ангел Кънчев”, Тел.: 082 888 685, E-mail: Irena@ecs.ru.acad.bg

Инж. Йордан Калмуков, Катедра “Компютърни системи и технологии”, Русенски университет “Ангел Кънчев”, Тел.: 082 888 827, E-mail: JKalmukov@gmail.com

Докладът е рецензиран.

Специализирани търсещи машини в Интернет пространството – основни принципи и архитектура

Ирена Вълова, Йордан Калмуков

Specialized Search Engines on the Internet - Basic Principles and Architecture

Irena Valona, Yordan Kalmukov

Abstract: *This paper describes the main principles and problems of building an automated tool for searching the web. Choosing the most appropriate search strategy seems to be the key point for achieving higher efficiency, denoted as the number of useful downloaded resources per time unit.*

Converting the directed weightless graph of WWW to a directed weighted one is absolutely necessary to find an adequate way for navigating through the web. The accuracy of weights' calculations directly determines how efficient the web crawler will be. Several ways of calculating the weight of vertices and edges are suggested and briefly described in this paper. The paper also proposes a common architecture of web spider designated to find and store images.

Key words: *web crawler, web spider, offline browser, automated tool for searching images on the web, searching strategy.*