УДК – 004
Econ Lit – C8

ПРИЛОЖЕНИЕ НА ТЕХНОЛОГИИТЕ ЗА ИЗВЛИЧАНЕ НА ЗНАНИЯ ОТ УЕБ В ЕЛЕКТРОННАТА ТЪРГОВИЯ

Доц. д-р Снежана Сълова

Резензенти: доц. д-р Тодорка Атанасова
доц. д-р Виолета Димитрова

Предговор

В съвременните условия на нарастваща конкуренция и икономическа нестабилност изискванията към оперативността и качеството на вземаните решения са много по-високи. Същевременно информацията, която трябва да бъде обработена за вземането на правилни и обосновани решения, става все повече. Непрекъснато нарастват и информационните източници в Интернет, които са ценен допълнителен източник на полезни знания. Всичко това изисква оптимално използване на наличните данни за получаване на повече знания и постигането на бизнес целите на организациите. Според традиционното годишно глобално проучване, проведено от Gartner[1] в началото на 2011 г., бизнес анализите са на пето място в списъка с десетте най-важни технологични приоритети на ИТ ръководителите[2].

Понастоящем съществуват много и разнообразни системи за бизнес анализи. Повечето от тях обаче разчитат на обработка и анализ на структурирани данни тези, които са натрупани в базите от данни. В много бизнес ситуации от съществено значение за бизнеса са и знанията, които могат да се извлекат от неструктурираната информация, например от интернет източници като: текст, хипервръзки, тагове, лог файлове и др. В сферата на електронната търговия, където основните бизнес дейности се осъществяват чрез използването на съвременните информационни, комуникационни технологии и Интернет от съществено значение, е анализът на уеб информация.

Цел на настоящата студия е да се направи изследване на процеса на извличане на знания от уеб ресурси и да се предложи модел на система за електронна търговия, в която са интегрирани дейностите по извличане на информация от интернет източници.

---

[1] Американски фирма, занимаваща се с анализи и изследвания в областта на информационните технологии.

[2] Интелигентните бизнес анализи – тенденции и перспективи. // CIO, № 3, 2011, (http://cio.bg/3715_inteligentnite_biznes_analizi__tendencii_i_perspektivi, 2.07.2011).

Едно от основните предизвикателства пред развитието на електронната търговия е приложението на интелигентните бизнес анализи, базирани на интернет ресурси. На тази основа **предмет** на настоящата студия е да се разработи нов метод за интегрирано използване на данните от фирмените бази от данни и от интернет източниците за осигуряване на по-задълбочени анализи.

Обект на изследване са системите за електронна търговия, като в студията се предлага модел на система за е-търговия, в която се осъществява интеграция на извличането на знания от различни източници.

I. Процес на извличане на знания от уеб източници

1.1. Същност на процеса на извличане на знания

Извличането на знания от данни (на английски терминът е Data Mining – DM) най-общо може да се определи като процес на изследване и откриване на скрити данни, знания, които преди това не са били известни, а са полезни за бизнеса чрез използване на техники и алгоритми от областта на изкуствения интелект[3]. За разлика от извличането на данни посредством заявки към база от данни, извличането на знания от данните генерира неочевидна, предварително неизвестна, но потенциално полезна информация. Този процес е възможен благодарение на анализа на големи обеми от данни. Едно от най-добрите определения за DM, което се е наложило в литературата, е на изследователите от Масачузетския технологичен институт (Massachusetts Institute of Technology – MIT). Те дефинират извличането на данни като „анализ на изследвани набори от данни (често големи), с цел откриване на неподозирани връзки и обобщаване на данните по нови начини, които са едновременно разбираеми и полезни за собственика на данните”[4].

За прецизността на бизнес анализите, основаващи се на извличането на знания, се налага тенденцията за търсене на нови и разнообразни източници на информация. Понастоящем може да се каже, че световната мрежа Интернет се обособи като най-богатия източник на лесно достъпна информация. В мрежата има голямо количество документи, данни аудио и видео файлове, от които чрез подходяща обработка биха могли да се извлекат нови и полезни за бизнеса знания. Познанието се генерира не само от съдържанието на самите страници, но също и от уникалните характеристики, структурата на уеб сайтовете и от информацията за достъпа до тях.

Основните характеристики, които определят Интернет като подходящ източник за нови знания, са:

- Големият обем и обхват информация. Световната мрежа Интернет с бурните си темпове на развитие обхваща все повече потребители, според последни данни броят им е достигнал 2,4 млрд. души[5]. Интернет служи за пренасяне на голямо разнообразие от информационни ресурси, които съдържат и голям обем полезна за икономическите единици информация.

[3] Барсегян, А. и др. Анализ даных и процессов. Санкт-Петербург, 2009, с. 68.

[4] Hand, D., B. Mannila and P. Smyth. Principles of Data Mining. MIT Press, 2001, http://mitpress.mit.edu/books/chapters/026208290Xchap1.pdf, (23.08.2011).

[5] Internet World Stats, http://www.internetworldstats.com/stats.htm, (8.12.2012)

Редица анализатори са доказали, че по аналогия със закона на Мур глобалната мрежа удвоява обема си на всеки 5,32 години и се очаква през 2012 г. да бъдат създадени 1,2 сетабайта уникална информация[6].

- Свързаността на информацията чрез използването на хипервръзки. Интернет служи за пренасяне на голямо разнообразие от информационни ресурси и услуги, най-известни сред които са взаимносвързаните хипертекстови документи. Известно е, че хипервръзките препращат към други документи, които имат връзка с разглеждания, освен това те могат да се проследяват автоматично от специални програми за търсене и индексация, наречени уеб spider или уеб crawler.
- Хетерогенният характер на информацията. В Интернет основните източници на информация са уеб сайтовете, които включват в своето съдържание информация под формата на текст, таблици, изображения, мултимедия (аудио, видео).
- Динамичността на информацията. Публикуваната информация в Интернет се обновява непрекъснато, появяват се актуални гледни точки за проблемите, което в много случаи се явява източник на нови знания.

Натрупването на все повече информация в уеб пространството е предпоставка за прилагането на познатите техники за извличане на знания от интернет източници. Извличането на знания от уеб ресурси се нарича Web Mining (WM). В литературата този термин се използва за първи път от Еционi през 1996 г. Той определя WM като процес на използване на DM техниките за изследване и извличане на информация от интернет документи и услуги[7]. По-късно понятието WM се обогатява, като към него се включват и техниките за изследване и анализ на данните за използваемостта на уеб ресурсите. Други автори дефинират процеса на Web Mining като „разкриване на закономерности, зависимости, шаблони в съдържанието, структурата и начина на използването на уеб документите, които показват как да се приложат DM методите и моделите върху уеб базирани данни”[8]. Като обобщение на цитираните определения може да се счита най-популярното в литературата становище за WM – „откриване и анализ на полезна информация от Световната мрежа”[9].

На базата на цитираните определения може да се дефинира, че извличането на знания от уеб представлява прилагане на технологиите за извличане на знания върху уеб ресурси – документи, хипервръзки, тагове, сървърни лог файлове и др.

---

[6] Кръстева, Н. Законът на Мур е в сила и за интернет ресурсите, http://cio.bg/4380_zakonat_na_mur_e_v_sila_i_za_internet_resursite, (10.01.2012).

[7] Etzioni, O. The World Wide Web: quagmire or gold mine? // Communications of the ACM, no. 11, 1996, pp. 65–68.

[8] Markov, Z. and D. Larosed. Data Mining the Web Uncovering Patterns in Web Content, Structure, and Usage. New Jersey: John Wiley & Sons, 2007.

[9] Cooley, R, B. Mobasher and J. Srivastava. Web Mining: Information and Pattern Discovery on the World Wide Web. (http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.27.3042&rep=rep1&type=pdf, 3.07.2011).

В зависимост от основните типове интернет ресурси, които се използват в процеса на WM, в литературата[10] са дефинирани три основни вида извличане на знания:

- Web content mining (WCM). Определя се като процес на извличане на полезни знания от съдържанието на уеб документите. Например, чрез WCM може: автоматично да се класифицират и групират (клъстеризират) уеб страници, според техните теми; да се изследва какво търсят потребителите в Интернет и да се откриват характерни шаблони за търсене на зададената група от хора; да се откриват модели в уеб страниците на базата на описанията на продуктите, съобщения с мнения от форуми и др. Освен това WCM се използва за подобряване намирането и филтрирането на информация, категоризиране на документи, намиране на подобни страници върху различни сървъри, идентифициране на теми, представени в различни уеб документи.
- Web structured mining (WSM). Процес на откриване на полезни знания в хипервръзките, чрез изследване на структурата от входящи и изходящи връзки (топологията) в уеб. Прилага се теорията на графите за анализиране на даден сайт и структурата му от връзки. Изследванията се групират в две направления: извличане на модели от хипервръзките на уеб и изследване на структурата на документа. Най-често се използва дървовидна структура за анализ и описание на таговете (HTML, XML) в уеб страницата. Например от връзките може да се открият важни уеб страници, общности от потребители, които споделят общи интереси и др.
- Web usage mining (WUM). Процес на откриване на потребителски модели на базата на данните за използване на интернет ресурси. Осигурява намиране на шаблони в потока от кликвания и асоциирани данни, събрани или генерирани като резултат от потребителската интерактивност с уеб сайтовете. За целите на WUM обикновено се анализират записите за логванията към сървърите и се получава информация за наборите от страници, обекти или ресурси, които често са достъпни от групи потребители с общи интереси. Получените знания са полезни за процеса на персонализация, за търсещите машини, за усъвършенстване на системите и др.

Процесът на откриване на знания в данните е постъпков. В литературните източници[11] броят на дефинираните етапи е различен в зависимост от това, дали са дадени в по-обобщен вид, или по-подробно. Ние се придържаме към следната декомпозиция на процеса на откриване на знания в данните:

---

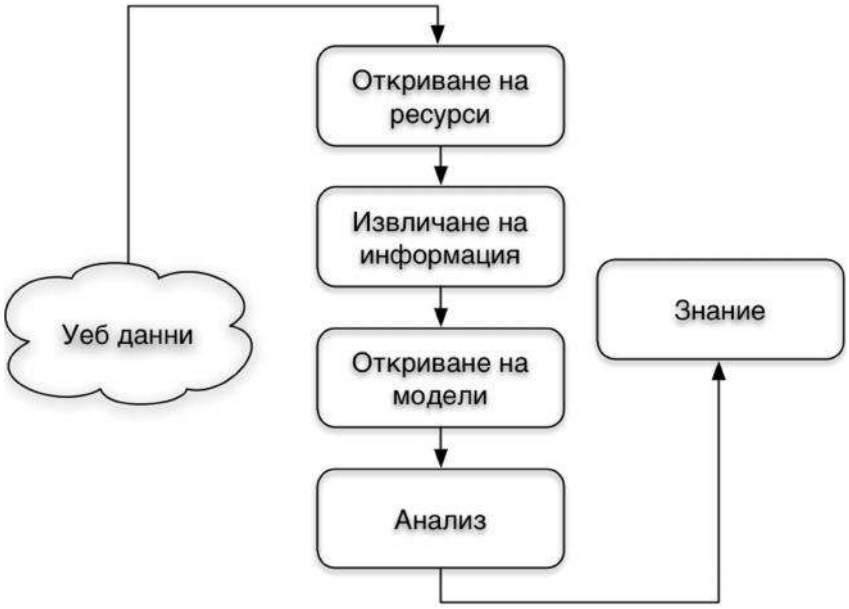
[10] Cooley, R. Mobasher, B. and Srivastave, J., Web Mining: Information and Pattern Discovery on the World Wide Web (http://maya.cs.depaul.edu/classes/ect584/papers/cms-tai.pdf, 9.12.2011). Markov, Z. and D. Larosed. Data Mining the Web Uncovering Patterns in Web Content, Structure, and Usage. New Jersey: John Wiley & Sons, 2007.

[11] Дефинирани според стандарта Cross Industry Standard Process for Data Mining - CRISP-DM (http://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining, 31.08.2011); Fayyad, U., G. Piatetsky-Shapiro and P. Smyth. From Data Mining to Knowledge Discovery in Databases (http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf, 27.08.2011).

1. Разбиране и формулировка на задачата за анализ.
2. Подготовка на данните за автоматизиран анализ, предварителна обработка.
3. Прилагане на методите на DM и построяване на модел.
4. Проверка на построените модели.
5. Интерпретация на резултатите.

Въпреки че извличането на данни от уеб източници води началото си от DM, необходимо е да се отбележи, че двата процеса не са идентични. Основната причина за това е, че традиционно извличането на информация използва структурирани данни, съхранявани в складове от данни, релационни таблици, електронни таблици или текстови данни, но в таблична форма. Уеб данните трябва първо да бъдат открити и след това да се сведат до вид, подходящ за анализ, тъй като те по своето същество са полу- или неструктурирани.

В литературата процесът на Web Mining също може се срещне декомпозиран по различен начин в зависимост от това, дали подзадачите са дадени подробно, или в по-обобщен вид. В изследването ние се придържаме към формулировка на етапите, направени от Косала и Блокийл[12] (фиг. 1):



Фиг. 1. Етапи на WM

[12] Kosala, R. and H. Blockeel. Web Mining Research: A Survey (http://facweb.cs.depaul.edu/mobasher/classes/ect584/papers/kosala.pdf, 4.07.2011).

1. Откриване на ресурси – търсене и извличане на информация от мрежата. Използват се търсещите машини, които обхождат милиони документи и индексират думите в уеб документите, запазвайки си индексите за последващо използване в други търсения.
2. Извличане на информация – автоматично избиране и предварителна обработка на специфична информация от определени уеб източници. Тази подготовка включва преобразуване на текстове чрез премахване на излишните думи, извличане на фрази и словосъчетания, съставяне на релации.
3. Откриване на общите модели на данните. Прилагат се методите на DM за машинно обучение, разпознаване на образи, статистически методи, невронни мрежи.
4. Анализ – анализиране на получените резултати. На този етап изключително значение има ролята на човека и неговата интерпретация на получените резултати.

При построяването на моделите за анализ на данните основните методи, които се използват, са:

- Класификация. В процеса на анализ често е необходимо да се определи към кой от някои известни класове се отнася изследваният обект. Например клиентите в един електронен магазин може да оценят и да определят „най-ценните" от тях. Класификацията се основава на данни за често повтарящи се покупки, похарчени суми и продължителност на интереса към фирмените продукти.
- Регресия. Задачата на регресията е сходна с класификацията и позволява да се определят по-известни характеристики на даден обект стойността на други параметри. За разлика от класификацията, при регресията множеството характеристики, по които се определя параметърът, са числа. Важно е да се отбележи, че основният проблем при решаването на задачи посредством класификация и регресия възниква от не доброто качество на изходните данни, в които може да има грешни данни, липсващи значения, атрибути от различни типове.
- Асоциация. Търсенето на асоциативни правила представлява намиране на общи зависимости между обектите или събитията. Намерените зависимости се представят като правила и могат да се използват за анализи и прогнози. Например от набора закупени стоки или услуги от клиентите може да се определят стоките или услугите, които често се закупуват заедно.
- Клъстеризация. Представлява търсенето на независими групи (клъстери) и техните характеристики в целия набор от данни. Групирането на данните позволява да се намали техния обем и улесни анализа. Например групирането може да се приложи към клиентите. Знае се, че потребителите са различни, имат различни потребности и изисквания към стоките и харесват различни начини за покупка. Това е предпоставка за сегментирането им и прилагането на диференцирани подходи при обслужването им.
- Последователност. Изследва се верижността на събитията. Например след покупката на жилище обикновено се закупува кухненско обзавеждане, кухненска техника и др.

Разгледаните основни методи за извличане на знания от уеб данни се основават на различни алгоритми. При класификацията се използват правила, базиращи се на алгоритмите: 1-R и Navie Bayes, ID3, C4.5 и SVM.

Алгоритъмът 1-R (1-rule) класифицира обектите, като създава по едно правило за всеки техен атрибут, след което се избира правилото с най-малко грешки като единствено. Той генерира дърво на решенията на едно ниво. В много от случаите обаче този алгоритъм е неприемлив, защото е необходимо при класификацията да се отчете значението на няколко атрибута на данните. Този недостатък на алгоритъма 1-R е преодолян в алгоритъма Naive Bayes. Неговият механизъм на работа се състои в изчисляване на условната вероятност за принадлежността на обекта към даден клас при наличието на известни значения на някои атрибути на данните. Алгоритъмът се базира на формулата на Бейс за изчисляване на условна вероятност за настъпване на дадено събитие, след като вече е известна част от информацията за него. За да бъде достоверна получената информация, чрез подхода на Бейс се изисква наличието на данни за изчисляване на пълните вероятности за характеристики на обектите.

Дърветата на решенията са метод за класифициране на данни чрез графи във формата на дърво. Те представляват структура от възли и връзки между тях. Дървото моделира процеса на вземане на решение относно класа, към който трябва да бъде отнесен класифицираният обект, чрез проверки на логически условия. Общият принцип на построяване на дърво на решенията се основана на изчисляване на условни вероятности и обособяването на подмножества. Този процес се повтаря рекурсивно, докато разделянето спре да добавя нови стойности и не може да има обособяване на следващи подгрупи. При построяването на дърветата на решения трябва добре да се избират характеристиките, по които ще се разделят множествата. Най-често използваните алгоритми са: мярка на „информационната полезност" (алгоритъм ID3), алгоритъмът C4.5 и индексът „Gini", реализиран в алгоритъма CART (Classification and Regression Tree).

Класификационният алгоритъм ID3 работи, като избира една характеристика за „най-добра" и по нея разделя множеството от обекти. Той изчислява показателят „информационна полезност" на базата на концепцията за ентропия на информацията[13] и избира характеристиката с най-голяма „информационна полезност", която образува възел при построяването на дървото. Възелът с ентропия 0 е лист и не се подлага на следващо разделяне. Възелът с ентропия по-голяма от 0 се подлага на разделяне, за да се класифицира съвкупността от обекти. Алгоритъмът C4.5 може да се разглежда като разширение на ID3. При него се изчислява „нормализирана информационна полезност" и по този начин се вземат само значимите възли при построяването на класификационното дърво.

Алгоритъмът CART е предназначен за построяване на двоично дърво на решенията. При него за построяването на всеки негов възел се използва правило, което дели съвкупността от обекти на две – една, за която правилото се изпълнява,

[13] Shannon, Cl. Prediction and entropy of printed English. // The Bell System Technical Journal, № 30 (http://www.princeton.edu/~wbialek/rome/refs/shannon_51.pdf, 12.01.2012).

и друга, за която то не се изпълнява. Този алгоритъм предоставя възможности за оценка на качеството на разклонението, за обработка на липсващи стойност и има механизъм за оптимално орязване на дърветата на решения.

За класификация могат да се използват и математически (линейни и нелинейни) функции, при които се изразява съотношението между зависими променливи и независими. Такива са статистическите методи и оптималният линеен класификатор Support Vector Machines (SVM), който разпределя обектите в класове, като графично ги представя в пространството.

За откриване на асоциативни правила най-разпространеният алгоритъм е Apriori. Той генерира често срещани подмножества от данните. Наборът от множествата се разширява едно по едно и процесът приключва, когато не се откриват по-нататъшни разширения. Apriori използва вид търсене в ширина (breadth-first search BFS)[14] търсене.

Множеството алгоритми за клъстеризация най-общо могат да се разделят на йерархични и нейерархични. Йерархичната клъстеризация от своя страна включва агломеративни и разделящи алгоритми. При агломеративните методи се извършват последователни сливания на единици и клъстери. Започва се с няколко клъстера, представляващи отделните единици, и след последователни сливания се достига до един клъстер, обединяващ всички единици. При разделящите методи подходът е обратен – започва се с един клъстер, който обединява всички единици, и след последователни разделяния се завършва с много клъстери, като всяка единица формира отделен клъстер. От разработените клъстерни методи най-често се използват йерархичните агломеративни методи. Резултатите при тях се представят чрез дървовидна диаграма, която графично показва йерархичната структура, получена от матрицата на сходство и правилата за обединение на клъстерите. Съществуват различни стратегии на обединение на обектите в клъстери и след това на самите клъстери. При образуване на клъстери във вид на „верига" най-често се използват методите на междугруповото свързване (Between-groups linkage) и на най-близкия съсед (Nearest neighbor).

От нейерархичните алгоритми най-разпространеният е на k-средните стойности (K-Means Cluster). В началото се избират k произволни изходни центрове (точки в пространството) и всички обекти се разделят на k групи в зависимост от това към кой център са най-близко. Разстоянието, близостта до центровете се определя по един от методите – Евклидово разстояние, Квадрат на евклидовото разстояние, Чебишево разстояние и др. След това се изчислява новият център на клъстера, на базата на средните значения на стойностите на обектите, вследствие на което се получава, че някои обекти вече попадат в друг клъстер. Процедурата се повтаря рекурсивно, докато центровете на клъстерите не престанат да се сменят. Този алгоритъм се използва при предварително зададен брой клъстери. Необходимостта потребителят да зададе броя на клъстерите като вход е един от основните му недостатъци.

[14] Алгоритъм за търсене, използва дървовидна структура, започва търсенето от основата и след това претърсва всички съседни възли.

За да се представи по-подробно същността и спецификите на извличането на знания, по-нататък се разглежда процеса по видове Web Mining.

1.2. Извличане на знания от съдържанието на уеб източници

Методите за извличане на информация от съдържанието на документите до голяма степен зависят от типа на ресурсите. Най-общо може да се дефинират две групи документи: неструктурирани и полуструктурирани. Към първата група се отнасят всички текстови документи, които нямат определена структура. Към втората група се причисляват документи, които като цяло имат структура, но съдържат и неструктурирани елементи, текст (такива са HTML, XML документите).

Търсенето на знания в текст е нетрадиционен процес на откриване на нови, потенциално полезни и разбираеми модели в неструктурирани текстови данни[15]. Процесът на анализ на текстовите документи преминава през следните стъпки:

1. Търсене на информация. Определя се какви документи трябва да се подложат на анализ. На този етап се използват търсещите машини (търсачки), които позволяват намиране на информация по зададени от потребителя термини (ключови думи). Може да се използват и специално разработени програми за търсене, които работят на базата на алгоритмите на някои от най-известните търсещи машини. Най-популярната търсеща машина е Google. Тя е с общо предназначение и предоставя възможност за търсене във всички области на знанието. Има и специфични търсачки, които търсят в бази от данни по определена тема или по специфичен отрасъл. По този начин увеличават възможността за намиране на релевантна информация в дадена област.

2. Предварителна обработка на документите. В текстовите документи има голямо количество думи и не е целесъобразно всички те да се подлагат на анализ, тъй като не всички са носители на полезна информация. Освен това някои от думите са синоними, имат едно и също значение и е добре да се разглеждат като един и същ термин. На този етап се премахват излишните, или т.н. стоп-думи (stop words) – думи, които са спомагателни, носят малко информация за съдържанието на документа, като например: „за", „по-горе", „през", „след", „така", „по-късно", „отново", „срещу" и др. Всички думи се преобразуват да бъдат изписани с малки букви. Освен това се прави представяне на думите в нормална форма, като се премахва множественото число, наставките и др. Като алтернатива преобразуването може да използва представянето на думата като множество от предварително определен брой символи. Например думата „маса" може да се представи като комбинации от 3 символа „_са", „мас", „аса", „ма_", като подчертаващото тире е предшествания символ или разделителя след думата. При този способ има по-малко граматически грешки, но се получават и много несмислени думи. В процеса на предварителната обработка се осъществява и преобразуване на символите, за да бъдат изписани само с малки букви.

3. Извличане на информация. На този етап се използват методите класификация, асоциация, клъстеризиране и др. Документите се разпределят в категории, из-

[15] Fayyad, U,, G. Piatetsky-Shapiro and P. Smyth. From Data Mining to Knowledge Discovery in Databases (http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf, 5.07.2011).

вършва се автоматично анотиране, извлича се най-съществената част от текста, като ключови понятия, имена, идентифицират се асоциативните отношения. Откриват се зависимости и модели, които се прилагат за анализ на информацията. В таблица 1 са дадени основните методи за извличане на информация от текстови документи.

4. Интерпретация на резултатите. На последния етап се визуализират резултатите, обикновено в графичен вид, което позволява да се извърши анализ и тълкуване.

Таблица 1

Алгоритми за извличане на съдържание от неструктурирани документи[16]

Автор	Алгоритъм	Начин за представяне на информацията	Приложение
Ahonen и др.	Eposode rules – асоциативни правила	колекция от думи	– за намиране на ключови думи и фрази – за откриване на граматически правила и словосъчетания
Billsus и Pazzani	– TFIDF – за определяне значението на думите в даден текст – Naive Bayes – базиран на теоремата на Бейз	колекция от думи	текстова класификация
Cohen	– Propositional rule based system – базира се на допускания – Индуктивно логическо програмиране	релация	текстова класификация
Dumais и др.	– TFIDF – за определяне значението на думите в даден текст – Дървета на решения – Naive Bayes – базиран на теоремата на Бейз – Мрежи на Бейс – Support Vector Machines	– колекция от думи – фрази	текстова категоризация
Feldman и Dagan	Относителна ентропия	концептуални категории	намиране на модели между текстовите данни
Feldman и др.	Асоциативни правила	терми	намиране на модели сред терми в текстовите данни
Frank и др.	Naive Bayes – базиран на теоремата на Бейз	фрази	намиране на фрази

[16] Kosala, R. and H. Blockeel. Web Mining Research: A Survey (http://facweb.cs.depaul.edu/mobasher/classes/ect584/papers/kosala.pdf, 4.07.2011).

Freitag и McCallum	Скрити модели на Марков	колекция от думи	обучаващи се модели за извличане
Hofmann	Статистически методи	колекция от думи	йерархично групиране
Honkela и др.	Self-Organizing Maps – самоорганизиращи се карти	колекция от думи	клъстеризация на текстове и документи
Junker и др.	Индуктивно логическо програмиране	релация	– категоризация на текст – обучаващи се модели за извличане
Kargupta и др.	– Неконтролируема йерархична класификация – Дървета на решения – Статистически методи	колекция от думи	– класификация на текст – йерархично групиране
Nahm и Mooney	Дървета на решения	колекция от думи	– предсказване на думи – зависимости
Nigam и др.	Максимална ентропия	колекция от думи	текстова класификация
Scott и Matwin	Rule based system - базира се на правила	– колекция от думи – фрази – синоними – хиперними[17]	текстова класификация
Soderland	Обучение по правла	– изречения – клузи	изучаване на правилата на изречението
Weiss и др.	Ускорени дървета на решения	колекция от думи	текстова категоризация
Wiener и др.	– Невронни мрежи – Логистична регресия	колекция от думи	текстова категоризация
Witten и др.	Компресиране на текст	– именувани единици	класификация и наименуване на същности
Yang и др.	– алгоритми за клъстеризация – на най-близкия съсед – дървета на решения	– колекция от думи – фрази	откриване на събития и проследяване

При анализ на полуструктурирани документи се използват по-развити средства за диференциране на текста, тъй като тези документи имат някаква структура. Алгоритмите за анализ използват HTML структурата и основните методи за извличане на знания като класификация, клъстеризация, асоциация. Основните от тях са обобщени и дадени в табл. 2.

[17] Хиперним – дума, която е по-обща от дадена дума.

Таблица 2

Алгоритми за извличане на съдържание от полуструктурирани документи[18]

Автор	Алгоритъм	Начин за представяне на информацията	Приложение
Craven и др.	– Modified Naive Bayes – модифициран алгоритъм на Бейз – Индуктивно логическо програмиране	релация и онтология	– хипертекстова класификация – извличане на връзки между уеб страници – изучаване на правилата за извличане
Crimmins и др.	Unsupervised and supervised classification algorithms - класификационни алгоритми	фрази, URL, метаинформация	– йерархична и графична класификация – клъстеризация
Fürnkranz	Обучение по правила	колекция от думи и информационни хипервръзки	– хипертекстова класификация
Joachims и др.	– TFIDF – за определяне значението на думите в даден текст – подпомагащо обучение	колекция от думи и информационни хипервръзки	– хипертекстово предсказване
Muslea и др.	Обучение по правила	колекция от думи, тагов и позиции на думите	изучаване на правилата за извличане
Shavlik и Eliassi-Rad	Невронни мрежи с подсилено обучение	колекция от думи и взаимоотношения	– хипертекстова класификация
Singh и др.	– Променени асоциативни правила – Класификационни алгоритми	понятия и наименования на обекти	намиране на шаблони в полуструктуриран текст
Soderland	Обучение по правила	изречения, фрази и наименование на обекти	изучаване на правилата за извличане

Съществуват и методи за извличане на уеб съдържание с цел формиране на бази от данни, което прави възможно последващото формиране на заявки към данните. Тези методи използват най-вече представяне на информацията във вида Object Exchange Model – OEM, който представлява модел за обмен на полуструктурирани данни между обектно ориентирани бази от данни. Информацията в OEM се пред-

[18] Kosala, R. and H. Blockeel. Web Mining Research: A Survey (http://facweb.cs.depaul.edu/mobasher/classes/ect584/papers/kosala.pdf, 4.07.2011).

ставя чрез графи с именувани ребра и възли, съответстващи на обектите. Всеки обект си има уникален идентификатор и няма ограничение за връзките между обектите. Основните използвани методи са дадени в таблица 3.

Таблица 3

Алгоритми за извличане на съдържание, базирани на метода OEM[19]

Автор	Алгоритъм	Начин за представяне на информацията	Приложение
Goldman и Widom	Частни алгоритми	OEM	търсене на DataGuides в полуструктурирани данни
Grumbach и Mecca	Частни алгоритми	стринг и релации	търсене на схеми в полуструктурирани данни
Nestorov и др.	Частни алгоритми	OEM	търсене на йерархии в полуструктурирани данни
Toivonen	Подобрени асоциативни правила	OEM	търсене на полезни подструктури в полуструктурирани данни
Wang и Liu	Подобрени асоциативни правила	OEM	търсене на повтарящи се подструктури в полуструктурирани данни
Zaiane и Han	Атрибутно-ориентирана индукция	релации	многослойни бази от данни

В много от случаите при работата с базите от данни се създават т.н. DataGuides[20]: кратки и точни структурни резюмета на полуструктурираните бази данни. DataGuides служат като динамични схеми. Те са полезни за разглеждане на структурата на базата от данни, формулиране на заявки, съхраняване на информация, като например статистически данни и примерни стойности. Позволяват се и заявки за оптимизация.

Чрез някои методи се извличат данни със задачата да създават многослойна база от данни, в която всеки слой се получава чрез обобщение на по-долните слоеве. За работа с този вид бази от данни се използва специален език за заявки и разработени специфични алгоритми за извличане на знания.

[19] Kosala, R. and H. Blockeel. Web Mining Research: A Survey (http://facweb.cs.depaul.edu/mobasher/classes/ect584/papers/kosala.pdf, 4.07.2011).

[20] Goldman, R. and J. Widom. DataGuides: Enabling Query Formulation and Optimization in Semistructured Databases, (http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.55.8594&rep=rep1&type=pdf, 5.07.2011).

1.3. Извличане на знания от структурата на уеб източници

Задачата на извличането на знания чрез изследване на структурата на интернет ресурсите е да се открие модел на връзките в мрежата. Web Structured Mining се базира на топологията на хипервръзките с или без описание на тези връзки. Откритият модел може да се използва за категоризиране на уеб страници и е полезен при генерирането на информация, например за сходството и отношенията между уеб сайтове.

През последните години структурата на уеб страниците широко се използва за анализ на важна информация. Изследванията на уеб структурата до голяма степен са повлияни от изследванията за социалните мрежи и анализите на цитиранията. Цитатите с връзки към уеб страниците са показател за голяма значимост. Най-общо връзките са 2 вида: in-links – за връзки, сочещи към страница, т.н. входящи връзки, и out-links – за връзки, сочещи към други страници, т.н. изходящи връзки. Обикновено, колкото е по-голям броят на връзките към дадена страница, за толкова по-полезна се счита тя. Освен това има и алгоритми за анализ, които изследват и текста, който описва връзките (т.н. anchor текст) и може да спомогне за по-качествен анализ.

Хипервръзките се използват за увеличаване на рейтинга на страницата, но те основно са пътищата, по които потребителите в Интернет се предвижват от една уеб страница към друга. Във връзка с това, популярността (или престижа) на дадена уеб страница може да се измери по отношение на това колко често един среден потребител посещава сайта. Уеб страниците с висок рейтинг са важни, защото са цитирани от много други страници и са посещавани често от потребителите.

Най-разпространените WSM алгоритми са: Page Rank и HITS (Hyperlink-Induced Topic Search). Чрез Page Rank се изпълнява класиране на страници с хипертекстови връзки между тях, използван в търсачката Google. Алгоритъмът е написан от Лари Пейдж (откъдето идва и името Пейдж-Ранг) и Сергей Брин, като част от проучване, проведено от Станфордския университет.

Основната идея на алгоритъма Page Rank е следната, ако например на една страница a има N връзки към други страници, между които е и страницата b, за потребителя, посетил страницата a, вероятността да посети b е $1/N$. Предлаганата схема за определяне на ценността на страниците, на базата на изходящите връзки, е количеството престиж, което страницата b получава от страницата a като $1/N$ от престижа й. Алгоритъмът за изчисление на Page Rank е рекурсивен. Уеб страницата е с висока оценка, ако е свързана към много други страници и те са с добър рейтинг.

Алгоритъмът HITS[21], известен още като Hubs and authorities, се използва за анализ на връзки. Разработен е от Джон Клайнбърг. Основната идея при него е първо да се фокусира на подходящите страници и след това да се изчисли значимостта им. В сравнение с Page Rank се работи с много по-малък граф. Започва се с търсене по ключова дума и след това се прави анализ на структурата на връзките.

[21] Manning, C., P. Raghavan, H. Schьtze, Hubs and Authorities (http://nlp.stanford.edu/IR-book/html/htmledition/hubs-and-authorities-1.html, 8.07.2011).

Процесът се нарича още извеждане на „същината на темата". При този алгоритъм се изчисляват две оценки, за хъб-страница, която сочи към много други, и за значимост- страница, към която сочат много други страници. В обобщение – страница с висок рейтинг е тази, от която има много изходящи връзки, както и страница, към която сочат много връзки.

Алгоритмите Page Rank и HITS се основават на свързаните страници и на това, че те принадлежат към една и съща тема. Ако обаче се увеличи броят на връзките, бързо се променя и темата. Увеличаването на страниците води до процес, известен като обобщение на темата. Значителното разширяване на набора от връзки увеличава броя на несвързаните страници и затова трябва да се избягва. Друга нежелана ситуация е, когато един тематичен набор от страници сочи към голям набор страници от друга тема и може да се получи отклонение от темата. Проблем може да възникне и при определянето на ранк и при плътно свързаните страници, които са разположени на един и същи сървър или се получава т.н. семейственост. Един лесен подход за избягване на семействеността е назначаването на тежест от 1/k на вътрешните страници на един сайт с общо k страници.

За преодоляване на проблемите с тези алгоритми за анализ е препоръчително да се използва не една система за определяне на значимостта, а да се приложат множество такива и да се претеглят техните стойности за изчисляване на крайния рейтинг на дадена страница.

1.4. Извличане на знания от информацията за ползваемостта на уеб източници

Процесът на извличане на знания на базата на интернет източници се фокусира върху техники, които помагат да се прогнозира поведението на потребителя, докато той взаимодейства с уеб. Базира се на данните за използването на уеб ресурсите, намиращи се в уеб клиента, прокси сървърите и сървърите.

Най-общо процесът на анализ може да се осъществи, като, първо, се извличат данни и се вмъкват в релационни таблици или директно да се използват данните от лог файловете, като се прилагат специални техники за извличане и анализ.

В научните изследвания процесът на WUM се разглежда в три етапа: предварителна обработка на данните, откриване на правила и модели и анализ на зависимости[22].

1. Предварителна обработка на данните. Осъществява се като се преобразува първичната информация в абстрактни данни, подходящи за по-нататъшните процеси. Основните информационни източници, достъпни за извличане, са: лог файловете, описанията на страниците, топологията на сайта, регистрите на потребителите и анкетите.

Може да се каже, че предварителната обработка е най-трудният етап от WUM. Основните проблеми на този етап произтичат от непълнотата на данните, особено ако се следи само по IP адрес. Някои типични проблеми са чрез един IP адрес могат да се осъществяват много сесии, тъй като някои интернет доставчици предо-

[22] Srivastava, J. et. al. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data http://portal.acm.org/citation.cfm?id=846188 (9.07.2011).

ставят набор от прокси сървъри за потребителите си. Възможно е доставчиците да използват средства за анонимност и да предоставят различни IP адреси за заявки от един потребител. Анализите се затрудняват, когато един и същ потребител осъществява достъпа си от различни компютърни устройства или от различни браузъри. За решаването на проблема с идентифицирането на потребителите се използват и анализират и кукитата[23]. След идентифицирането на потребителите се осъществява разделяне на клик потока на сесии. Този процес също понякога е труден за осъществяване, тъй като не е лесно да се установи в кой момент потребителят е напуснал уеб сайта. За разделянето на сесии обикновено се използва интервал от 30 минути[24]. Освен това анализът на сесиите се затруднява и от някои сървъри, които поддържат променливи за всяка сесия и се налага да се осъществява достъп до сървъра на съдържание, за да се разбере каква заявка е зададена и какво се визуализира при потребителите.

В процеса на WUM е необходимо и осъществяване на предварителната обработка на съдържанието. Това е процес на конвертиране на текст, изображения, скриптове, мултимедия във форма, подходяща за извличане на модели от уеб източници. Приложението на технологиите на Data Mining към съдържанието на уеб сайтовете може да се използва като филтър за входа или изхода на алгоритмите за откриване на модели. Например приложението на алгоритмите за класификация може да ограничи откриването на модели само за разглеждани страници, които са за точно определен продукт. Класификацията или клъстеризирането на съдържанието на страниците помага за определяне на целта на посещението. При предварителната обработка текстовете от страниците се представят като вектори от думи. Изображенията и мултимедийните представяния се заместват от ключови думи. Както беше споменато и при описанието на WCM, съдържанието на статичните уеб страници се обработва сравнително лесно в сравнение с динамичните страници, при които има използване на бази от данни, персонализация и се визуализират различни изгледи. За да се извърши обработка на динамичните страници, се използват http заявки, програми тип роботи за търсачки, скриптове за достъп до бази от данни.

Етапът на предварителната обработка за WUM изисква и обработка на структурата – хипервръзките между страниците. Структурата може да се извлече и обработи по същия начин като съдържанието. Тук отново по-трудни за обработка са динамичните страници, които имат различна структура при всяка сървърна сесия.

2. Откриване на правила и модели. Това е следващият етап от WUM. При него се прилагат методи и алгоритми от областта на статистиката, извличането на данни, машинното обучение, разпознаването на образи с цел откриването на модели в процеса на уеб анализиране. Най-често използваните техники за извличане на знания за посетителите са статистическите. Чрез различни видове статистически анализи върху сесийните данни може да се намерят средното време за разглеждане на страницата, най-често посещаваните страници, дължината на навигационния път. На базата на анализ на грешките може да се открият опитите за неауторизи-

[23] HTTP кукита е пакет информация, изпратен от уеб сървър към интернет браузър и след това връщан от браузъра всеки път, когато той получи достъп до този сървър.

[24] Catledge, L. and J. Pitkow. Characterizing Browsing Strategies in the World-Wide Web (http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.103.4010&rep=rep1&type=pdf, 12.07.2011).

ран достъп. Получените знания от статистическите анализи са полезни за подобряване на производителността, сигурността или се използват за маркетингови цели.

Други методи за извличане на знания, които могат да се приложат на този етап, са асоциацията, клъстеризацията, класификацията, анализ на последователностите и моделиране на зависимости. Генерирането на асоциации се използва за откриване на връзки между страници, които се отварят често заедно. Откритите зависимости се прилагат за бизнес цели, а също са в помощ на уеб дизайнерите при преструктуриране на сайтовете. Друго полезно приложение на асоциациите е за предвиждане на интересите и предварително извличане на документи, за да се предотврати забавянето при зареждане на страниците.

Методът за групирането на елементите с подобни характеристики – клъстеринг, също се използва при WUM. Групират се потребители със сходни шаблони за разглеждане на уеб страници. Подобно знание е изключително полезно за извършване на демографски анализ на потребителите, който да послужи за пазарни сегментации и осигуряване на персонализирано уеб съдържание. Клъстерите може да се използват и за откриване на група страници със свързано съдържание. Тази информация е полезна за търсещите машини, които предлагат страници в зависимост от заявките и миналата история на потребителите.

В WUM се използва и класификацията – разпределяне на обектите в предварително дефинирани класове. В уеб пространството целта е да се разработи профил на потребители, принадлежащи към определен клас. Класификацията на сървърните логове може да доведе до получаването на интересни зависимости.

За изследване на потребителските сесии е приложима и техниката за откриване на последователностите, която прави опит за определяне на вътрешносесийни шаблони, при които присъствието на набор от елементи е последвано от друг елемент за определен период от време. Анализът на последователностите е удобен за прогнозиране на потребителското поведение и за откриване на тенденции.

Откриването на зависимости е друг полезен начин за създаване на модели за WUM. Неговото приложение е най-вече за създаване на модели на поведение и е потенциално полезно за предсказване на бъдеща потребителска консумация.

3. Анализ на моделите. Това е последният етап от WUM. Целта на този процес е да се филтрират намерените модели и правила и да се запазят само тези, които са значими и интересни за изследването. Използва се обикновено графично представяне на данните или различно оцветяване на стойностите, за да се открояят цялостните модели и тенденции в данните.

Направеното изследване на процеса на извличане на знания от уеб източници позволява да се направи изводът, че съществуват множество методи за извличане на информация от съдържанието, структурата и от данните за използването на уеб ресурсите. Тенденцията за непрекъснатото увеличаване на използваемостта на световната мрежа Интернет води до нарастване на значението и полезността на съдържащата се в уеб източници информация. Извличането на знание от уеб е от съществено значение за целите на анализите в редица сфери на бизнеса и затова предлагаме интегриране и използване на WM анализите в системите за електронната търговия.

II. Извличане на знания от уеб източници в електронната търговия

2.1. Области на приложение

Електронната търговия се определя като „използването на Интернет за осъществяване на търговия”[25] или най-общо представлява извършване на бизнес транзакции, които се подпомагат от съвременните информационни и конуникационни технологии. Е-търговията вече е реална част от ежедневието на много хора по света. Данните от проучванията показват непрекъснато увеличаване на приходите от онлайн търговията. Оценките са през 2011 г. те да нараснат с 18,9% и да станат 680 млрд. долара[26].

Клиентите пазаруват онлайн поради удобния и лесен начин на достъп до каталозите с предлаганите стоки и услуги. В интернет магазините е възможно предлагането на голям асортимент от стоки, който физически не може да бъде изложен или разгледан в обикновен магазин, даже и в склад. Освен това тези стоки и услуги са непрекъснато достъпни за всички географски места, а потребителите са улеснени при покупката, защото не губят време за посещаване на магазините и получават покупките направо в домовете си. Друго предимство на Интернет търговията е, че търговците имат възможност за динамично определяне на цените и бързо адаптиране към променящите се бизнес условия, а клиентите изразходват значително помалко време за изследване на пазара и бързо достигане до желания продукт или услуга. Електронната търговия предоставя добри възможности за представяне на фирмите и позволява и на малките фирми да бъдат конкурентоспособни и с успех да работят на международния пазар.

Една от основните причини за бързото развитие на електронната търговия е възможността за персонализация и индивидуално обслужване на клиентите. Подоброто управление на връзките с клиентите с цел запазване на съществуващите клиенти и спечелването на нови, по-гъвкаво ценообразуване, избор на рационални дистрибуторски услуги и осъществяването на оптимални и навременни доставки са от съществено значение за онлайн търговията. За осъществяването на тези дейности е необходимо широко да се използват бизнес анализите. Според нас основните области на приложение на интелигентните бизнес анализи за е-търговията са:

- персонализиране на клиенткото обслужване;
- моделиране на потребителското поведение;
- идентифициране на свързани стоки;
- групиране и оценка на клиентите;
- планиране на продуктовия асортимент;
- прогнозиране на пазарното търсене;
- наблюдение на удовлетвореността на клиентите;
- изчисляване нивата на поддържаните запаси;
- избор на рационална схема за доставка;

[25] Fingar, P., H. Kumar and T. Sharma. Enterprise E-Commerce. Tampa: Meghan-Kiffer Press, USA, 2000.
[26] Rao, L., P. Morgan. Global E-Commerce Revenue To Grow By 19 Percent In 2011 To $680B (http://techcrunch.com/2011/01/03/j-p-morgan-global-e-commerce-revenue-to-grow-by-19-percent-in-2011-to-680b/,10.07.2011).

- избор на подходящ доставчик за стоките;
- изчисляване обема и периодичността на доставките;
- групиране на поръчки.

За извършване на по-прецизни анализи в областта на електронната търговия считаме, че е необходимо да се използва и наличната уеб информация. Събраните данни за клиентите при регистрация или чрез въпросници невинаги са носители на достоверна и пълна информация. Много от клиентите от липса на достатъчно мотивация или от съображения за сигурност избират да не попълват своите данни.

Уеб анализите, макар и да не са изключително прецизни, са носители на много и полезна информация за потребителските посещения, за интересите на потребителите, за тяхното поведение. Те имат голямо значение за персонализирането на услугите и за насочването на потребителите към информация, която би била максимално полезна за тях. Анализите на уеб ресурси подпомагат разбирането как се използват уеб сайтовете за търговия и услуги и какво търсят най-често потребителите. Те могат да са и в помощ на маркетинговите специалисти за изучаване и сегментиране на пазара. Освен това този вид анализи се използват и за изучаване на конкуренцията, което е важен фактор за успешен бизнес.

Като основни области на приложение на уеб анализите в електронната търговия може да се обособят следните:

- **Идентификация и обособяване на групи за маркетингови цели.** На основата на данните за използването на сайтовете и на базата на информацията от уеб съдържанието се извлича знание за клиентите и се разширяват познанията за основните клиенти на съответния електронен магазин. Чрез алгоритмите за клъстеризация се идентифицират групи за маркетингови цели, извличат се знания за тези групи, на които по-късно се предлагат специални оферти, съобразени с техните интереси. Например може клиентите, закупили дадена стока да се обособят в една група и на базата на анализ на използваемостта на сайта от тях да се установи, че те са разглеждали една и съща друга стока. Това би могло да послужи за основание стоката, към която се проявява интерес да се предложи на съответната група клиенти на преференциални цени.
- **Оценяване на клиентите и контрагентите.** Осъществява се класификация и клъстеризация на клиентите и контрагентите на базата на анализ на посещенията на сайта и на записаните мнения във форумите. На идентифицираните като „най-ценни” и лоялни клиенти на електронния магазин се предлагат промоционални оферти, а на контрагентите – преференциални условия за сътрудничество. Например на базата на мненията на клиентите може да се идентифицират тези, които по-често записват положителни мнения за стоките и съответно да се класифицират към групата на „най-ценните” клиенти.
- **Подпомагане на персоналното обслужване.** Извличането на информация за предпочитанията и нагласите на клиентите към отделни продукти, услуги дава възможност за индивидуален подход при обслужването на всеки потребител. Следенето на предпочитанията и нуждите на клиентите позволява прилагането на специално отношение към всеки. Целият набор от дейнос-

ти, свързани с грижата и вниманието към клиентите, гарантира тяхното задържане. Анализите на сървърните лог файлове също са източник на знания за потребителските предпочитания, например разглежданите страници помагат да се натрупа информация за това какво харесва съответният потребител.

- **Препоръки за нови покупки.** Процесът на генериране на препоръки е важен за търговците на дребно и доставчици на услуги. Клиенти, към които се насочват подходящи и навременни фирмени препоръки, купуват повече и са по-лоялни. Най-добър пример в това отношение е онлайн магазинът Amazon.com, при които се извеждат множество препоръки от типа „Потребителите, закупили тази стока, са закупили и ...", „Посетителите, които са разгледали тази стока, са се заинтересували и от ...", „Специално за вас...". За всички тези препоръки са необходими не само данните за покупките, но и извлечените данни от съдържанието и използваемостта на сайта.
- **Откриване на последователности от свързани събития и стоки, които се закупуват заедно.** Използва се методът на асоцииране за откриването на събития, които се случват заедно, и стоки, които се купуват заедно. На базата на анализ на добавените от клиентите в потребителската кошница стоки може да се открият продукти, които външно изглеждат несвързани, но често се купуват заедно. Информацията от сървърните сесии може да бъде източник на откриването на закономерности в поведението на потребителите. Анализът на последователностите може да помогне за планиране на стоковите запаси, например ако се клиентът закупи фотоапарат и разглежда аксесоари за него, това е признак, че те са му необходими и скоро може да си ги купи.
- **Сегментиране на пазара и генериране на информация за търсенето.** Анализират се посещенията, изследва се значимостта на уеб страниците и се извлича информация за потребителските интереси и търсения, с цел да се удовлетворяват клиентските желания. Разделянето на сегменти има голямо значение за маркетинговата стратегия на фирмата, защото благодарение на нея е възможно да се осигури предлагане на стоки и услуги, които в максимална степен отговарят на очакванията и предпочитанията на потребителите.
- **Анализ на успеха на рекламните кампании.** Осъществява се на базата на информацията от потребителските форуми и мнения, изследване на рейтинга на страниците и данните за посещаемостта на рекламните и промоционални сайтове. От данните за достъпа до рекламните сайтове може да се извлече знание за поведението на клиентите и да се направят оценки, дали те са се повлияли от рекламите и са добавили в своите пазарски кошници стоки, които се продават промоционално.
- **Планиране на продуктовия асортимент и набора от услуги.** Анализира се покупателното поведение на клиентите на фирмата. На базата на записаните мнения и интереса на клиентите към стоките се получават знания за това какви стоки е най-добре да се предлагат в електронния магазин, което може да спомогне за оптимизация на стоковия асортимент.

- **Откриване на тенденции и прогнозиране.** Прогнозирането на търсените количествата и видове стоки на базата на анализ на мненията и данните за направените покупки може да доведе до значително повишаване на приходите от продажбите.
- **Откриване на изключения и контрол на риска.** Оценяване, идентификация и намиране на редки случаи, които се различават съществено от нормите. Намира приложение най-често за борба с престъпленията с кредитни карти и идентифициране на опити за неоторизиран достъп до системата.
- **Подобряване на системата за е-търговия.** Чрез анализ на информацията за използването на уеб ресурсите се получава обратна връзка от потребителите и се преценява дали структурата и съдържанието на сайта са добри. Освен това информацията за достъпа до системата може да покаже дали има опити за неавторизиран достъп и налагат ли се подобрения по сигурността.

За да се открои по-добре използването на различните видове уеб анализи в е-търговията, сме разработили функционална матрица на тяхното приложение. В нея са дадени дейностите и чрез какъв тип WM може да се осъществи анализът (таблица 4). Таблицата е създадена на базата на описаните в литературните източници възможности, които предоставя съответния вид WM.

Таблица 4 показва, че най-приложими в сферата на е-търговията са аналите на използваемостта на уеб ресурсите. Те могат да дадат отговор на въпроси като: „Коя е най-често посещаваната уеб страница от сайта?", „В какъв ред обикновено се преглеждат страниците?", „Кои други сайтове посещават потребителите?", „По колко уеб страници разглежда типичният потребител?", „Колко време е обичайният престой на потребителите на сайта?", „След разглеждането на коя страница потребителите най-често напускат сайта?". Анализирайки получените отговори на тези въпроси, може да се изследва на поведението на даден потребител или съвкупното поведение на голям брой потребители в рамките на определен период от време и да се използва полученото ново знание за стимулиране на продажбите и увеличаване на приходите от търговската дейност.

Широко приложение намира и извличането на знания от съдържанието на уеб ресурсите. Основно при този вид WM се събира, категоризира, организира налична информация за стоките и потребителите. Извлеченото ново знание допълва съществуващата в базите от данни информация и предоставя възможност за осъществяване на по-точни анализи на покупателното поведение на клиентите и на факторите, които оказват влияние при вземането на решение за покупка. Освен това получените познания за клиентите позволяват на фирмите да прилагат редица гъвкави инструменти и средства на маркетинга за повишаване на обема продажбите и печалбата на фирмата. Изследването на връзките между страниците и значимостта им е също източник на полезна информация и може да послужи за извличане на важни за търговската дейност зависимости, както и за предприемане на дейности за повишаване на сигурността и ефективността на системите за е-търговия.

Таблица 4

Приложение на WM анализите в електронната търговия

Дейност	Видове уеб анализи		
	Web Content Mining	Web Structure Mining	Web Usage Mining
Идентификация и обособяване на групи за маркетингови цели	+	+	+
Оценяване на клиентите и контрагентите	+	-	+
Подпомагане на персоналното обслужване	+	+	+
Препоръки за нови покупки	+	+	+
Откриване на последователности от свързани събития и стоки, които се закупуват заедно	+	-	+
Сегментиране на пазара	-	+	+
Анализ на рекламните кампании	+	+	+
Планиране на продуктовия асортимент и набора от услуги	+	-	+
Откриване на тенденции и прогнози	+	+	+
Откриване на изключения и контрол на риска	-	+	+
Подобряване на системата за е-търговия	+	+	+

От представената матрица може да се направи изводът, че уеб анализите са от съществено значение за електронната търговия. Интернет ресурсите и извличането на нови знания от тях помага да се разбере „защо" са се случили съответните събития, да се открият тенденции, да се направи асоциация с подобни събития. Предоставя се възможност за осъществяване на така важните за е-търговията – индивидуално обслужване на всеки клиент, по-добро управление на връзките с клиентите, гъвкаво ценообразуване, рационални доставки и съставяне на прогнози за бъдещо развитие.

2.2. Интеграция на извличането на знания в системите за е-търговия

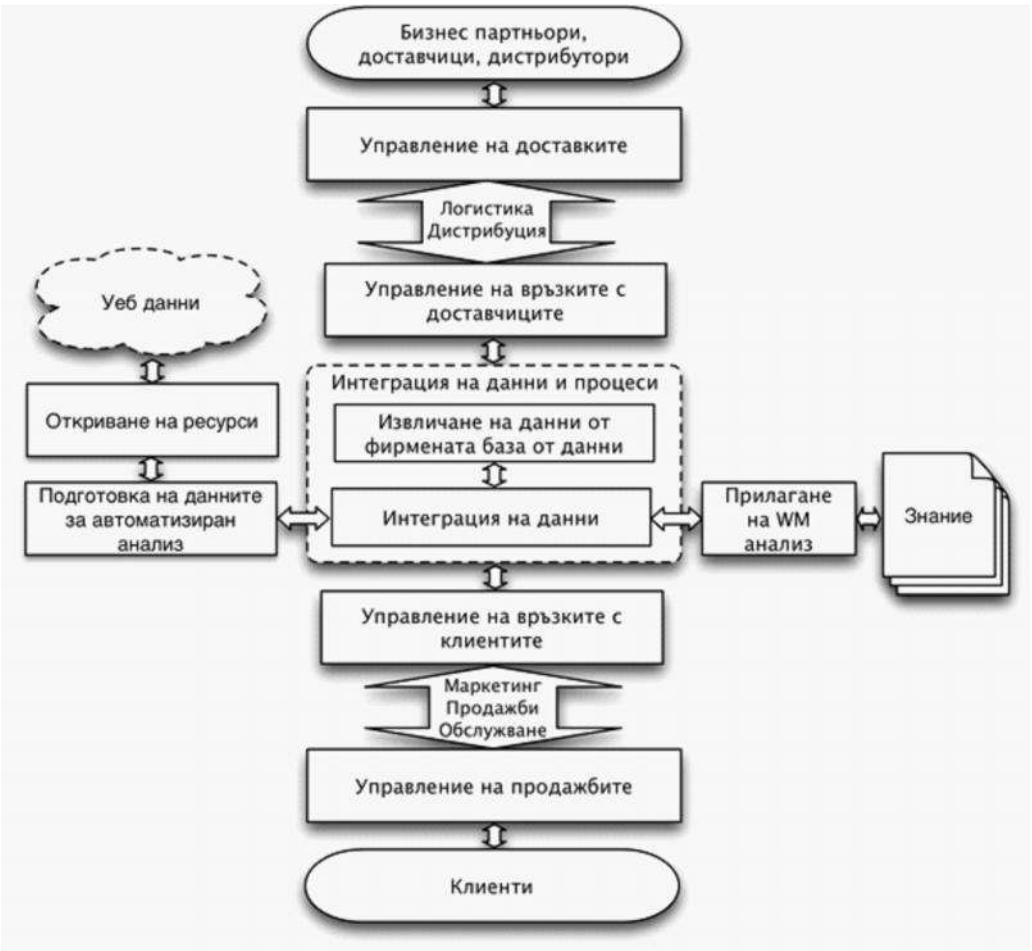
Електронната търговия обхваща процесите на купуване и продаване на стоки и услуги посредством компютърни мрежи, комуникиращи си чрез глобалната мрежа Интернет. Основните бизнес дейности, които автоматизират системите за е-търговия, са:

- управление на връзките с клиентите;
- управление на продажбите;
- управление на връзките с доставчиците;
- управление на доставките.

В процеса на осъществяване на всяка една от тези дейности е необходимо използването на бизнес анализите за вземане правилни решения и максимизиране на ефективността на онлайн бизнеса.

Системите за електронна търговия представляват уеб базирани приложения и като такива те са достъпни през Интернет посредством уеб браузър. Те могат да работят като самостоятелни софтуерни приложения или да са интегрирани в някаква степен или цялостно към фирмените компютърни информационни системи. При големите фирми обикновено онлайн магазините се интегрират с останалите фирмени приложения и се използва обща фирмена база от данни. При тях все по-често се използва и технологията на складовете от данни, която позволява съхраняване и проследяване на историческа, архивна информация, консолидация на големи обеми данни от различни подсистеми и предоставя средства за анализи и прогнози. Всичко това позволява на приложенията за е-търговия да се възползват от широкия спектър от възможности на фирмените системи за бизнес анализи.

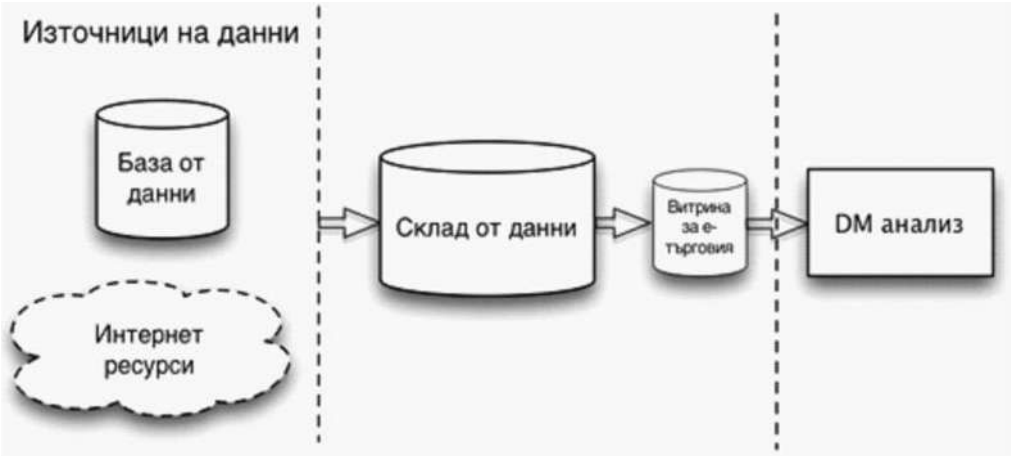
Системите за електронна търговия на по-малките фирми от своя страна разчитат или на вградените средства, или на справките и анализите от допълнително инсталиран софтуер. В двата разгледани случая обаче, бизнес анализите се осъществяват въз основа на натрупаните данни в базата от данни. Все по-голямата конкуренция между онлайн продавачите през последните години води до търсене на повече възможности за спечелване на нови клиенти, за по-добро обслужване на съществуващите клиенти, за предлагане на необходимите и качествени стоки и услуги. Всичко това създава необходимостта от търсене на нови източници на информация, чрез които да се допълнят и обогатят знанията за бизнеса. Ето защо предлагаме за целите на бизнес анализите в сферата на електронната търговия да се осъществява интегрирано използване както на полезните знания, които се съдържат в големите обеми информация, натрупана в базите от данни, така и от обработката на неструктурирана информация, извличана предимно от Интернет източници. Модел на система за електронна търговия, в която са интегрирани и възможностите за осъществяване на анализи на базата на информация от интернет източници, е показан на фиг. 2.



Фиг. 2. Интеграция на WM анализа в система за е-търговия

Считаме, че интеграцията на анализите, базирани на извличането на данни от интернет ресурси в система за е-търговия, може да помогне за получаване на повече знания за развитието на онлайн бизнеса, тъй като се използва и съществуващият огромен обем информация в световната мрежа. Необходимостта от прецизни анализи и достоверни прогнози, базирани на повече информационни източници, е основната причина за потребността от интеграция. В много от случаите извлечената от базата от данни информация е непълна и за да се извърши по-детайлен анализ, се налага използване и на WM технологията. Например, за да се направи анализ за интересите на определен клиент, може да се изследват неговите отзиви, коментари, които да се извлекат от съдържанието на уеб страниците и същевременно да се вземат за анализ и данните за направените от него покупки. По този начин получената информация ще е по-точна и пълна, а това ще позволи осъществяване и на по-задълбочени анализи, които ще допринесат за по-добро персонално обслужване и генериране на по-точни препоръки за съответния клиент.

Ще разгледаме реализацията на модела от технологична гледна точка и в зависимост от информационното осигуряване на фирмите. Традиционно при големите фирми като Amazon и HP системите за електронна търговия са източник на много данни. Те работят с обща фирмена база от данни, в която обикновено има възможност за съхранение и на историческа и архивна информация. За осъществяването на по-ефективно обобщаване и анализиране на големите обеми от данни, които се натрупват, обикновено тези фирми използват технологията на складовете от данни. Това им позволява прилагането на OLAP и Data Mining анализите върху съхранените данни. Данните, постъпващи във фирмения склад от данни, се извличат от различни фирмени системи и от външни източници. За целите на изследването се разглеждат постъпващите данни от системата за е-търговия и тези, които са извлечени от интернет източници. Предлага се създаването на специална витрина от данни (Data Marts)[27], която да съдържа данни, извлечени от склада, и данни, постъпващи от интернет източници (вж. фиг. 3). Обобщаването на извлечените от склада знания с тези, получени при WM анализа, ще позволи генерирането на нови за бизнеса модели и зависимости, които са източник на полезна информация.



Фиг. 3. Извличане на знания при фирми, използващи склад от данни

При по-малките фирми също може да се използва предлаганият на фиг. 2 модел. При тях ще се обобщават извлечените от уеб данни с данни, извлечени директно от базата от данни, тъй като обикновено тези фирми нямат изграден склад от данни, а създаването на такъв е скъпо струваща инвестиция. Най-подходящо е данните при тях да се записват в табличен вид, за да може след това да се обработват по-лесно (фиг. 4).

[27] Използват се за съхраняване на подмножества от обобщените данни, дефинирани по определено свойство.

Фиг. 4. Извличане на знания с използване на данни директно от базата от данни

Етапите на процеса извличане на знание, които съответстват на предложения модел за интеграция на анализите на уеб информация в система за електронна търговия, са показани на фиг. 5. Процесът е декомпозиран по следния начин:

1. Откриване на интернет ресурси. Необходимо е предварително много добре да се разбере задачата за анализ и след това с помощта на търсещите машини да се намерят различни видове ресурси, текстови файлове, данни за използваемостта (лог файлове) и др.
2. Предварителната обработка на данните. На този етап се третират липсващите стойности, като за целта може да се игнорират съответните ред или колона от таблицата, да се направи попълване на тези стойности на базата на експертни оценки или средна стойност. Осъществява се също идентификация и обработка на силно отличаващи се и грешни стойности и се съгласуват данните, тъй като може да има несъответствие в кодирането и наименуването на определени категории данни.
3. Обобщаване на данните. Извлечените данни от интернет ресурсите при необходимост се допълват с данни от базата от данни или склада от данни. Създават се обобщени таблици.
4. Прилагане на средствата за DM анализ. Прилагат се различни техники и алгоритми от областта на машинното обучение, разпознаването на образи, статистиката, невронните мрежи за откриване на смислени корелации.
5. Извеждане на резултатна информация. Осъществява се под формата на справки, извеждане на двумерни и тримерни графики, система от показатели и табла.

По същество етапите са стандартните за извличане на знания от уеб ресурси, но е предвидено в етапа, в който автоматично се откриват общите модели на данните, да се осъществява обединяване с допълнително извлечени данни от фирмените бази от данни. Тези допълнителни данни при необходимост ще допълват извлечените от уеб данни, за да може при прилагането на избрания и подходящ метод

Фиг. 5. Етапи в процеса на извличане на знание от интернет източници

за анализ да се получи по-пълна информация. По този начин получените модели ще са валидни за по-голямо количество данни и резултатите от анализа ще са по надеждни.

За да се покаже приложението на интеграцията на WM анализите в система за е-търговия, ще разгледаме един пример. Целта е да се съберат повече данни за клиентите, за да се осъществи персонализираното им обслужване при тяхното следващото посещение в магазина. Използват се съхранените в базата от данни данни за предишните покупки на клиента. От таблица 5 се вижда, че клиент с номер 101 е закупил 2 стоки: номер 6 – фотоапарат Canon 550D, и номер 9 – обектив Cannon EF-S 18–55 mm F 3.5–5.6 IS. На базата на анализ на клик потока на клиента може да изведат страниците, които той е посещавал. Установява се, че освен закупените стоки той е разглеждал страниците на Filter Canon UV, фотоапарат Canon EOS 1100D и фотоапарат Canon PowerShot. Чрез анализ на съдържанието на посетените страници се установява, че клиентът е записал рейтинг на продукта Canon PowerShot и е оставил положително мнение за него.

Таблица 5

Данни за покупките на клиенти от електронен магазин

Номер на транзакция	Номер на клиент	Номер стока
1	101	6
2	101	9
3	105	12
4	204	3
...	...	...

Чрез обобщаване на извлечените чрез технологиите DM и WM данни може да се направи изводът, че изследваният клиент се интересува от фото техника с марката Cannon, тъй като и данните за покупките му и анализът на неструктурираната информация го показват. Това може да послужи за основание при следващо посещение на клиента в електронния магазин да му се предложат продукти от същата марка, както и преференциални оферти за тях.

2.3. Софтуерни средства

От технологична гледна точка за осъществяването на извличането на полезни знания не съществува универсален алгоритъм за намиране на решение на всички задачи. Търсенето и извличането на информация се осъществява с използването на различни програми, които от своя страна прилагат комбинация от представените в т. 1 алгоритми за извличане на знания. С цел интеграция между различните софтуерни решения има разработени стандарти за обмен на получените Data Mining модели. Такива са: CRISP-DM (Cross Industry Standard Process for Data Mining), PMML (Predictive Model Markup Language), CWM-DM (Common Warehouse Model for Data Mining) и др. Стандартът PMML понастоящем е най-използваният[28]. Той е XML-базиран. Разработен е от Data Mining Group (DMG). Позволява споделянето на модели между PMML-съвместими системи. Чрез него може да се дефинират независими от производителите модели, които да се използват от различни приложения. Улеснява се значително обмена на модели между приложенията. Основните предимства на PMML са, че:

- осигурява платформена независимост на извлечените знания;
- улеснява използването на DM модели от други приложения;
- позволява запис на данни във формат, подходящ за анализ чрез много приложения.

[28] Барсегян, А. и др. Анализ даных и процессов. Санкт-Петербург, 2009, с. 259.

За извличане на полезни знания от уеб източници се използват софтуерните средства за Data Mining, които имат и възможности за анализ на уеб информация, както и специално разработени програми за конкретен вид уеб анализ. Всички програмни продукти са създадени на базата на достиженията на статистическия софтуер, софтуера за разпознаване на образи, методите на изкуствения интелект и теорията на базите от данни.

В зависимост от лицензите за използването му специализираният софтуер за извличане на знания най-общо се дели на две основни категории: лицензионен софтуер и софтуер с отворен код. Лицензионният софтуер е собственост на дадена фирма, която има всички права по интелектуалната собственост и използването му. Този вид софтуер обикновено е платен. При софтуера с отвореният код се предоставя възможност за свободно използване на първичния код на програмите при спазване на определени в дефиницията на този вид софтуер критерии[29,30]. През последните години към софтуера с отворен код се проявява все по-голям интерес. Той се използва главно, защото много от организациите го адаптират, за да останат независими от цените и лицензионните политики на големите софтуерни фирми. Софтуерът с отворен код предлага сравнително добро качество, висока надеждност, голяма гъвкавост и възможност за развитие. Той е предпочитан от много потребители, защото се разпространява безплатно и има широка подкрепа на специалистите, занимаващи се с този вид разработки.

Програмните продукти за извличане на знания могат да работят самостоятелно или да бъдат компонент от цялостна бизнес интелигентна система. Често за целите на анализите се използват паралелно по няколко средства. В таблици 6 и 7 са представени основните специализирани софтуерни средства за Data Mining анализи, които имат и възможности за извличане на знания от уеб източници.

Таблица 6

Комерсиален софтуер за извличане на знания

Софтуер	Описание
Oracle Data Mining (ODM)	Част от пакета Oracle Database 11g. За WM основно се използва Oracle Text, който може да извършва лингвистичен анализ на документи, както и търсене на текст с помощта на различни стратегии, включително търсене по ключови думи, контекст заявки, чрез булеви операции. Има средства за класификация, групиране и визуализация на информацията.
SAS Enterprise Miner	Java базиран софтуер, разработка е на американската фирма SAS Institute. Има широк набор от възможности за подготовка на данните, обощаване и изличане на модели и визуализация. Използва специален модул Link Analysis за анализ на log файлове. Софтуерът на SA се използва основно за откриване на финансови измами, анализ с цел намаляване на риска, за съставяне на прогнози и за маркетингови цели.

[29] Илиева, С., В. Лилов, И. Манова. Подходи и методи за реализация на софтуерни решения. София: Св. Климент Охридски, 2010, с.32.

[30] The Open Source Definition (http://opensource.org/docs/osd, 16.03.2011).

SPSS Modeler и SPSS Text Analytics	Софтуер на фирмата IBM. Първоначално SPSS Modeler се е казвал SPSS Clementine. През 2009 г. е преименуван на SPSS Modeler. Има много добри интерфейсни средства и позволява извличане на знания от неструктурирана информация, чрез създаване на точни модели и прогнози. SPSS Text Analytics е специализирано средство за обработка на естествен език. Софтуерът е много добре интегриран с бизнес приложенията на IBM, което улеснява извличането на знания.
STATISTICA Data Miner	Пакет от софтуерни продукти за анализи и решения, на фирмата StatSoft. Софтуерът включва инструменти за анализ на масиви от данни, за управление и визуализация на данни, както и инструменти за групиране, класификация, моделиране и извличане на данни. Освен това предоставя и допълнителни възможности, достъпни благодарение на интеграцията му с програмния език R и неговата програмна среда.

Таблица 7

Софтуер с отворен код за извличане на знания

Софтуер	Описание
Carrot2	Приложение за Text Mining и клъстеризиране. Основно се използва за клъстеризиране на малки колекции от документи. Написано е на Java.
ELKI (Environment for DeveLoping KDD-Applications Supported by Index-Structures)	Използва се за клъстерен анализи и за откриване на изключения. Университетска разработка, написана на Java.
KNIME (Konstanz Information Miner)	Интегрира различни компоненти за машинно обучение и извличане на данни. Използва се основно във фармацевтиката, управлението на връзките с клиентите и за финансови анализи. KNIME е написан на Java и се базира на Eclipse, неговия механизъм за добавяне на разширения и плъгини, които предоставят допълнителна функционалност.
Orange	Компонентно базиран софтуер за машинно обучение и извличане на данни. Включва обширен набор от компоненти за предварителна обработка на данни, извличане на модели и техники за изследване. Разработка на университета в Любляна. Написан е на C++ и Python.
R	Език за програмиране със среда за разработка за статистически изследвания, извличане на данни и графично представяне. Синтаксисът му е базиран на Scheme. R предлага голямо разнообразие от статистически и графични техники, включително линейно и нелинейно моделиране, класически статистически тестове, анализ на времеви серии, класификация, групиране, и др. Разработен е от Ross Ihaka и Robert Gentleman от университета в Аукланд, Нова Зенландия.
RapidMiner, преди YALE (Yet Another Learning Environment), RapidNet	RapidMiner осигурява среда за машинно обучение и извличане на данни. Софтуерът е Java базиран и е разработи в университета в Дортмунд. Разполага с интерактивен, графичен потребителски интерфейс и средства за предварителна обработка, класификация, регресия, клъстеризация, асоциация и за графично представяне на моделите и резултатите. RapidNet е модул за анализ на структурата, който използва графичния потребителски интерфейс и средствата за визуализиране на RapidMiner.

UIMA (Unstructured Information Management Architecture)	Платформа за анализ на неструктурирано съдържание като текст, аудио и видео, първоначално разработена от IBM.
Weka (Waikato Environment for Knowledge Analysis)	Набор от алгоритми за изпълнение на задачи по извличане на данни. Weka съдържа инструменти за предварителна обработка на данните, класификация, клъстеризация, асоциативни правила и визуализация. Софтуерът е разработен на Java в университета Вайкато, Нова Зенландия.

Всички представени в таблиците софтуерни средства могат да се определят като програмни пакети за извличане на знания с общо предназначение, тъй като се базират на множество методи и алгоритми и са приложими за различни видове анализи. Основните предимства на този вид софтуер са:

- разполага с графичен потребителски интерфейс;
- има възможност за работа с различни типове данни;
- позволява работа с големи обеми от данни;
- предоставя средства за добро визуално представяне на резултатите;
- използва стандартизирани и общодостъпни комуникационни структури, интерфейси, протоколи и технологии.

Освен представените в таблици 6 и 7 софтуерни средства за извличане на знания с общо предназначение са разработени и някои специализирани програми за конкретен вид анализ на уеб информация като:

- AeroText – софтуерен пакет за text mining, които се използва за анализ на съдържание;
- Analog (на Dr. Stephen Turner) – за анализ на лог файловете на уеб сървъра;
- ClickTracks – за анализ на поведението, на базата на потребителските кликвания;
- Nihuo Web Log Analyzer – за анализ на посещенията на уеб сайтовете;
- Surf Pattern Visual Analyzer – за анализ на навигационата стуктура и хипервръзките;
- Textlyser – за анализ на текст;
- WebTrends – инструменти за анализ на посещаемостта на сайтовете;
- VISITaTOR – за клъстерен анализ на посетители на даден сайт; htminer – за анализ на лог файлове и др.

Тези програмни средства обаче са предназначени само за конкретен вид анализ, базират се на отделни методи и алгоритми за извличане на знания и работят с точно определен вид данни, поради което и се използват сравнително по-малко.

Проучванията сочат, че най-разпространени са програмните продукти за извличане на данни с общо предназаначение – R, STATISTICA, SPSS Modeler, Rapid

Miner[31]. Анализът на добрите практики показва, че основните причини, поради които са се наложили тези софтуерни продукти, са:

- всички използват различни методи и алгоритми за търсене на скрити зависимости, тъй като при извличането на знания не съществува един универсален алгоритъм за намиране на решение на всички задачи;
- разполагат със средства за визуализиране на намерения модел, след като системата е генерирала решение;
- осигуряват възможност за интеграция с други системи, използват стандарта PMML, което позволява споделянето на модели;
- предоставят възможност за извличане на данни от различен тип източници;
- предлагат версии за различни платформи Windows, Linux и Mac OS X;
- поддържат се и се усъвършенстват непрекъснато.

Цитираните най-разпространени софтуерни средства за извличане на знания са известни повече като софтуер за DM, но имат и добри възможности за работа с интернет ресурси.

Програмният продукт STATISTICA има отделни модули STATISTICA Text Miner за обработка на неструктурирана текстова информация и STATISTICA Sequence, Association and Link Analysis (SAL) – за анализ на асоциативни правила и структурата от връзки. SPSS Modeler разполага с функционалности за анализ на неструктуриран текст и за анализ на структурата от връзки. Благодарение на много добрата му интеграция с пакета R може значително да се разширява обхвата на решаваните задачи.

Rapid Miner разполага с интерактивен графичен потребителски интерфейс и средства за анализ на връзките, анализ на неструктуриран текст и лог файлове. Процесът на извличане на знания се реализира чрез дървета с оператори. Rapid Miner съдържа инструменти за предварителна обработка, класификация, регресия, клъстеризация, асоциативно обучение по правила и за графично представяне на получените резултати. Като основни източници на данни може да използва складове от данни, бази от данни, като SQL Server, Oracle, Access, MySQL, а също и CSV, Excel файлове.

Програмните продукти за извличане на знания с общо предназаначение използват широк набор от разгледаните в първа точка алгоритми и методи за WM, което ги прави и подходящи за използване при извличане на знания от Интернет източници. Допълнително тяхно предимство са и добрите им възможности за визуализация на получените резултати.

Трябва да се отбележи, че изборът на конкретно софтуерно решение зависи от множество фактори, като функционалност, възможности за интеграция със съще-

[31] 2010 Data Miner Survey (http://www.rexeranalytics.com/Data-Miner-Survey-Results-2010.html, 31.08.2011); Goebel, M. and Le Gruenwald. A survey of data mining and knowledge discovery software tools (http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.130.1456&rep=rep1&type=pdf, 31.08.2011); Data mining. From Wikipedia, the free encyclopedia (http://en.wikipedia.org/wiki/Data_mining, 31.08.2011); Herschel, G. Magic Quadrant for Customer Data-Mining Applications, Gartner Inc. (http://www.asiaminer.com.tw/Assets/Download/Gartner_Magic_Quadrant.pdf, 7.01.2012); Haughton, D. at al. A Review of Software Packages for Data Mining. // The American Statistician, Vol. 57, № 4, pp. 290–309. Most Popular Data Mining Software, (http://www.the-data-mine.com/bin/view/Software/MostPopularDataMiningSoftware, 31.08.2011).

ствуващия софтуер и цена. Няма универсално и най-добро решение за извличане на знания в областта на електронната търговия. Трябва да се изследва приложимостта на всеки отделен пакет за конкретния вид бизнес и организация. Необходима е комплексна предварителна оценка на възможностите на софтуера за успешна интеграция със съществуващите системи. Наложително е точно и ясно да се дефинират и изискванията спрямо софтуера за извличане на знания и ограниченията по отношение на време, персонал, парични средства и др. Едва след това чрез сравнение може да се избере софтуерен пакет, който да бъде полезен на фирмата.

Обикновено големите фирми, които вече са инвестирали средства за закупуване на системи за управление на бизнеса и са интегрирали електронните си магазини, с тези системи избират комерсиален софтуер за извличане на знания, който най-лесно би се интегрирал със съществуващите им системи. Подходящи са STATISTICA и IBM SPSS Modeler.

За начинаещ и неголям бизнес е добре да се предпочете по-евтино, но достатъчно функционално решение, като продуктите R и Rapid Miner. Изхождайки от резултатите от последните проучвания, които показват, че се увеличава използването на софтуера за извличане на данни с отворен код, може да се каже, че тези софтуерни решения са достатъчно добра алтернатива на комерсиалния софтуер. Често софтуерът с отворен код е напълно безплатен, което позволява преглеждане на неговия програмен код и дори модифицирането му за нуждите на конкретната организация. Освен това и много от успешно работещите в момента приложения за електронна търговия се базират именно на софтуера с отворен лиценз и тяхната интеграция с продукти като R и Rapid Miner е улеснена.

Заключение

Анализите на инфрмация от интернет източници намират приложение в области, където на са достатъчни само статистическите и аналитичните методи и изградените чрез тях модели. В направеното изследване се разглежда приложението на интелигентните бизнес анализи, базирани на извличането на знания от уеб ресурси за подобряване обслужването на клиентите, за откриване на тенденции, за асоциации с подобни събития и за съставяне на прогнози за бъдещото развитие. Представя се модел на система за е-търговия с интеграция на дейностите по извличане на знания от уеб източници и се предлага подход за реализацията му за различни видове фирми.

Считаме, че все по-широкото разпространение на е-търговията налага търсенето на нови възможности за усъвършенстване на електронните магазини. Понастоящем повечето функциониращи онлайн обекти са с много добър дизайн и функционалност, имат подробна и добре структурирана информация за продуктите и възможност за извеждане на оперативни отчети. За успешната им работа обаче, в динамичната пазарна среда, считаме, че е необходимо да се прилагат и интелигентните бизнес анализи, базирани на информация от интернет ресурси. Те са носители на много и полезна информация за потребителските посещения, за интересите на купувачите, за тяхното поведение и имат голямо значение за персонализирането на услугите и за насочването на потребителите към информация, която би била мак-

симално полезна за тях. Получените нови знания биха могли да спомогнат за подобряване и усъвършенстване на процесите по управление на взаимоотношенията с клиентите и на други маркетингови дейности, а знаем, че за фирмите, осъществяващи е-търговия, това е от изключително значение, тъй като при тях от лоялността на клиентите до голяма степен зависят и приходите от продажбите. Освен това този вид анализи могат да се използват от мениджърите за създаване на по-успешни бизнес стратегии, базирани на получените по-задълбочени и прецизни анализи и прогнози. Използването на извличането на нови знания от интернет ресурси би могло да бъде важно конкурентно предимство за фирмите, занимаващи се с електронна търговия, защото като цяло допринася за подобряване на маркетинговата и търговската им дейности.

Използвана литература

1. Атанасова, Т. Интелигентни компютърни системи. Варна: Наука и икономика, 2011.
2. Барсегян, А. и др. Анализ даных и процессов. Санкт-Петербург: БХВ-Петербург, 2009.
3. Илиева, С., В. Лилов, И. Манова. Подходи и методи за реализация на софтуерни решения. София: Св. Климент Охридски, 2010.
4. Кашева, М. и др. Изследване на бизнес интелигентните системи за малки и средни предприятия. Варна: Наука и икономика, 2011.
5. Кисимов, В., К. Стефанова. Принципи и методи за изграждане на разпределени бизнес интелигентни екосистеми. // Годишник на УНСС, том. 1, 2010.
6. Кръстева, Н. Законът на Мур е в сила и за Интернет ресурсите, (http://cio.bg/4380_zakonat_na_mur_e_v_sila_i_za_internet_resursite, 10.01.2012).
7. Пейков, В. Data Mining за оптималното управление на взаимоотношенията с клиентите. // CIO, № 12, 2009, Ай Си Ти Медиа.
8. Интелигентните бизнес анализи – тенденции и перспективи. // CIO, № 3, 2011 (http://cio.bg/3715_inteligentnite_biznes_analizi__tendencii_i_perspektivi, 2.07.2011).
9. BI проектът – стъпка към по- високо аналитично и бизнес равнище. // CIO, №10, 2008.
10. A. Guazzelli, M. Zeller, W. Chen, and G. Williams. PMML: An Open Standard for Sharing Models. // The R Journal, Vol. 1, № 1, May 2009.
11. Ansari, S. at. al. E-Commerce and Data Mining: Architecture and Challenges (http://ai.stanford.edu/~ronnyk/integratingEcom.pdf, 27.09.2011).
12. Berry, M. and G. Linoff. Data Mining Techniques For Marketing, Sales, and Customer Relationship Management. Indianapolis: Wiley, 2004.
13. Buddhinath G. and D Derry. A Simple Enhancement to One Rule Classification (http://www.buddhinath.net/OtherLinks/Documents/Improved%20OneR%20Algorithm.pdf, 7.11.2011).
14. Catledge, L. and J. Pitkow. Characterizing Browsing Strategies in the World-Wide Web (http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.103.4010&rep=rep1&type=pdf, 12.07.2011).
15. Chakrabarti, S, et al. Mining the Link Structure of the World Wide Web (http://www.cs.cornell.edu/home/kleinber/ieee99-web.pdf, 12.01.2012).

16. Cho, Y., J. Kim. Application of Web usage mining and product taxonomy to collaborative recommendations in e-commerce. Expert Systems with Applications 26, 2004, pp. 233–246.
17. Cooley, R, B. Mobasher and J. Srivastava. Web Mining: Information and Pattern Discovery on the World Wide Web (http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.27.3042&rep=rep1&type=pdf, 3.07.2011).
18. Cooley, R., P. Tan and J. Srivastava. Discovery of Interesting Usage Patterns from Web Data (http://www.cs.umn.edu/research/technical_reports.php?page=report&report_id=99-022, 12.01.2012).
19. Curtis, G. and Cobham, D. Business Information Systems. Financial Times Management, 2005.
20. Etzioni, O. The World Wide Web: quagmire or gold mine? // Communications of the ACM, № 11, 1996, pp. 65–68.
21. Fayyad, U, G. Piatetsky-Shapiro and P. Smyth. From Data Mining to Knowledge Discovery in Databases (http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf, 5.07.2011).
22. Fingar, P., H. Kumar, T. Sharma. Enterprise E-Commerce. Tampa: Meghan-Kiffer Press, USA, 2000.
23. Goebel, M., Le Gruenwald. A survey of data mining and knowledge discovery software tools (http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.130.1456&rep=rep1&type=pdf, 31.08.2011);
24. Goldman, R. and J. Widom. DataGuides: Enabling Query Formulation and Optimization in Semistructured Databases (http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.55.8594&rep=rep1&type=pdf, 5.07.2011).
25. Hand, D., H. Mannila, P. Smyth. Principles of Data Mining. MIT Press, 2001 (http://mitpress.mit.edu/books/chapters/026208290Xchap1.pdf, 23.08.2011).
26. Herschel, G. Magic Quadrant for Customer Data-Mining Applications, Gartner Inc. (http://www.asiaminer.com.tw/Assets/Download/Gartner_Magic_Quadrant.pdf, 7.01.2012).
27. Kadav, A., J. Kawale, P. Mitra. Data Mining Standards (http://www.datamininggrid.org/wdat/works/att/standard01.content.08439.pdf, 12.09.2011).
28. Kantardzic, M. Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons, 2003.
29. Kohavi, R. Mining E-Commerce Data: The Good, the Bad, and the Ugly, (http://robotics.stanford.edu/~ronnyk/pakdd2001.pdf, 12.09.2011).
30. Kosala, R. and H. Blockeel. Web Mining Research: A Survey (http://facweb.cs.depaul.edu/mobasher/classes/ect584/papers/kosala.pdf, 4.07.2011).
31. Larose, D. Data mining methods and models. Hoboken: John Wiley & Sons, 2006.
32. Manning, C., P. Raghavan, H. Schьtze. Hubs and Authorities (http://nlp.stanford.edu/IR-book/html/htmledition/hubs-and-authorities-1.html, 8.07.2011).
33. Markov, Z., D. Larosed. Data Mining the Web Uncovering Patterns in Web Content, Structure, and Usage. New Jersey: John Wiley & Sons, 2007.
34. Mikut, R., M. Reischl. Data Mining Tools. // Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Vol. 1, № 5, September/October 2011, pp. 431–443.

35. Mladenic, D, M. Grobelnic. Feature selection for unbalanced class distribution and Naive Bayes (http://kzi.polsl.pl/~jbiesiada/prace_magisterskie/TrappLudynia/html/Bibliografia/mladenic99feature.pdf, 5.07.2011).
36. Mobasher, B., R. Cooley and J. Srivastava. Automatic Personalization Based on Web Usage Mining. // Magazine Communications of the ACM, Vol. 43, №8, 2000.
37. Rao, L., J. Morgan. Global E-Commerce Revenue To Grow By 19 Percent In 2011 To $680B (http://techcrunch.com/2011/01/03/j-p-morgan-global-e-commerce-revenue-to-grow-by-19-percent-in-2011-to-680b/, 10.07.2011).
38. Shannon, Cl. Prediction and entropy of printed English. // The Bell System Technical Journal, № 30 (http://www.princeton.edu/~wbialek/rome/refs/shannon_51.pdf, 12.01.2012).
39. Srivastava, J. et. al. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data (http://portal.acm.org/citation.cfm?id=846188, 9.07.2011).
40. Wu, X., V. Kumar. The Top Ten Algorithms in Data Mining. Chapman and Hall/CRC, 2009.
41. Zupan, B., J. Demsar. Open-Source Tools for Data Mining (http://eprints.fri.uni-lj.si/893/1/2008-OpenSourceDataMining.pdf, 9.09.2011).
42. 2010 Data Miner Survey (http://www.rexeranalytics.com/Data-Miner-Survey-Results-2010.html, 31.08.2011).
43. Data mining. From Wikipedia, the free encyclopedia (http://en.wikipedia.org/wiki/Data_mining, 31.08.2011).
44. Most Popular Data Mining Software (http://www.the-data-mine.com/bin/view/Software/MostPopularDataMiningSoftware, 31.08.2011).
45. RapidMiner (http://www.rapidminer.com, 20.11.2011).

APPLYING THE TECHNOLOGIES FOR MINING KNOWLEDGE OUT OF THE WEB IN E-COMMERCE

Assoc. Prof. Dr Snezhana Salova

Abstract

Of great significance for business is knowledge that can be derived from unstructured information contained in Internet sources, such as text, hyperlinks, tags, log files, etc.

The paper presents a study of the process of extraction of useful knowledge from web resources. There is proposed a model of a system of e-commerce, in which there are integrated the activities of mining knowledge out of Internet sources. There has been developed a functional matrix of application of the various kinds of web mining for the sphere of e-commerce. There are also defined the required software tools for the realization of the presented model.

ANWENDUNG DER TECHNIKEN ZUR INFORMATIONSGEWINNUNG IM INTERNET IM ELEKTRONISCHEN HANDEL

Doz. Dr. Snezhana Salova

Zusammenfassung

Das Wissen, das die Subjekte der Wirtschaft aus den nicht-strukturierten Informationen im Internet gewinnen können, ist von wesentlicher Bedeutung für die Wirtschaft. Es handelt sich um Texte, Hyperlinks, Tags, Log-Dateien u.a.

In der Studie wird der Vorgang der Gewinnung nützlichen Wissens aus Web-Ressourcen untersucht. Ein Model eines Systems elektronischen Handels wird vorgeschlagen, in dem die Tätigkeiten zur Gewinnung von Wissen aus dem Internet mit dem Handel selbst integriert sind. Eine funktionelle Matrix zum Einsatz verschiedener Arten von Web Mining für den Bereich des elektronischen Handels wird dargestellt. Software-Mittel zur Verwirklichung des dargestellten Models werden definiert.

ПРИМЕНЕНИЕ ТЕХНОЛОГИЙ ИЗВЛЕЧЕНИЯ ЗНАНИЙ ИЗ ВЕБ СТРАНИЦ В ЭЛЕКТРОННОЙ ТОРГОВЛЕ

Доц. д-р Снежана Сылова

Резюме

Существенное значение для бизнеса приобретают знания, которые можно извлечь из неструктурированной информации, содержащейся в интернет источниках, таких, как тексты, гиперссылки, теги, лог файлы и пр.

В разработке проведено исследование процесса извлечения полезных знаний из веб ресурсов. Предложена модель системы электронной торговли, в которой интегрирована деятельность по извлечению знаний из интернет источников. Разработана функциональная матрица применения различных видов Web Mining в сфере электронной торговли. Определены софтуерные средства, необходимые для реализации представленной модели.

С ъ д ъ р ж а н и е