

ИЗЛИЧАНЕ НА ОБЕКТИВНИ ЗНАНИЯ ОТ ИНТЕРНЕТ

Ивайло Пенев, Пламен Пенев

Резюме. Статията предлага подход и помощни средства за отделяне на важната информация по интересуваша ни тема в Интернет. Описва се демонстрационна обобщаващо търсеща система „Справочник“, съдържаща елементи на изкуствен интелект. Реализиран е прост механизъм за получаване на дедуктивни определения. Изгражда се честотен речник и предметен указател. Запазването на ключовите думи в база знания позволява извеждане на умозаклучения по аналогия.

Keywords: frequency dictionary, index, knowledge base, classifier of concepts, conclusions, analogy.

1. Увод.

В заобикалящата ни действителност съществува безкрайно множество от обекти и свойства, които се отразяват в нашето съзнание във вид на понятия. Специфично за учебния процес в средното училище е, че в началото ученикът буквално се сблъсква с утвърдената система от термини и понятия. Тя липсва в предишния му опит и памет, възниква и се формира с усилие на мисълта и волята. Запознаване с определението е първа стъпка за усвояване на едно или друго понятие. Методическите условия, при които протича процесът на усвояване, са следните:

1. Активна мисловна дейност от страна на учениците. Репродуктивният метод, ориентиран на механично запомняне, не дава желаните резултати. Много по-ефективни са евристичните методи, осъществявани в интерактивна среда.

2. Осъзнаване на съществените и несъществените признаци на понятието. Отделянето на несъществените признаци предпазва от грешно обобщение, когато ученикът приема несъщественния признак за ориентир при извеждане на дадено определение.

3. Включване на новото понятие в системата на изучените до този момент понятия.

4. Нагледно изучаване на понятието. Особеното е, че обект на изучаване е думата, словосъчетанието, изречението. Следователно наред с таблиците, схемите, изображенията на предмети застава и езиковият материал.

В условията на повсеместна електронизация на учебния процес авторите на настоящото изследване си поставиха за цел разработване на автоматизирани средства за подпомагане на дейността на учителя при изясняване на смисъла на новите понятия. С други думи, възникна идеята за разработване на обобщаващо търсеща информационна система (ОТИС), която да извлича знания от публикуваните в Интернет документи.

Знанията в глобалната мрежа са представени на естествен език. Те имат следните предимства:

- Бърза достъпност.
- Наредба по посещаемост – на първи места се поставят най-очакваните сайтове.
- Актуалност на съдържание – за да се задържат на предни места, сайтовете се обновяват непрекъснато.

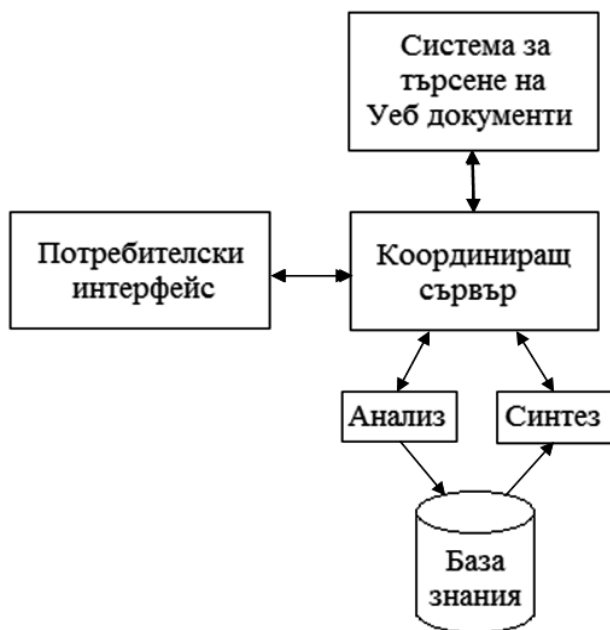
От друга страна, Интернет, поради анонимността на авторите, е предмет на постоянни критики относно достоверността на представяните факти. Дори компетентната Уикипедия не прави изключение. Как да извлечем интересувашите ни съществени думи и изрази от „разводнена“ тема? Традиционните търсещи инструменти (Google, Yahoo и други) не обработват смисъл. Те са базирани върху статистическото индексирание на предоставените текстове в зависимост от популярността на ключовите думи. Обстоятелството, че документът е бил предпочитан от повече потребители, не изключва случайния фактор и често резултатите са незадоволителни. Налага се повторно търсене, връщане назад и т.н. Клъстеризация, пренареждане, групиране по теми, които извършват допълнително някои метатърсещи системи (vivisimo, clusty), водят понякога до получаване на по-добра прецизност, но в други случаи се наблюдава влошаване на резултатите.

Повишаване на качеството на търсене е възможно само чрез извършване на предварителен синтактичен и семантичен анализ на текста. В последно време се появиха търсачки, изцяло базирани на семантични технологии. Типичен представител е Nakia. Инструментът допуска свобода на заявката, отличава се с бързо действие, формулира отговори, извежда определения, но с пълна функционалност предимно на английски език. Основно тази причина накара авторите да потърсят собствен път за решаване на проблема. Задачата се улеснява от факта, че се изисква обработка по строго определени правила при уточнена заявка. Необходима е олекотена, но ефективна търсеща система. Резерви за разработване на такава виждаме в техниките на метатърсене. Демонстрираната тук ОТИС „Справочник“ (<http://volov.iservice.bg:88/glagol/index.htm>) е метатърсачка, ползваща резултатите от ИТС (Yahoo, Google, Alexa) и извършваща допълнителен семантичен анализ.

2. Инструментални средства и състав на системата. Използваните инструментални средства не се различават от общоприетите стандартни:

- Мрежови компоненти на Delphi за автоматично подаване на заявки към ИТС и първична обработка на резултатите.
- Функции за работа със символни низове в JavaScript и Php при провеждане на семантичен анализ.
- Език за формулиране на заявки SQL за поддържане на релационна база данни.
- За координиране на работата на отделните компоненти спомага многорамковата организация на потребителския интерфейс.

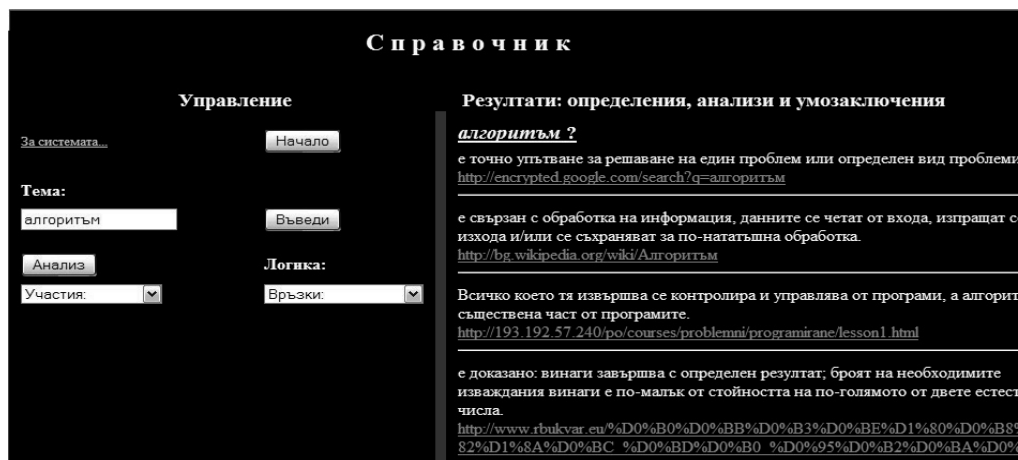
Сървърната платформа е Apache.



Фиг. 1. Основни компоненти на обобщаващо търсещата информационна система

3. Отделяне на класификационни определения. Концептуално-семантичните отношения (КСО) в общоупотребим текст и моделите за тяхното разпознаване са описани в [1]. Тук се разглежда най-често срещаното отношение: род-вид (правило $\langle X \rangle \text{ E } \langle Y \rangle$). Програмнотърсената връзка лесно се открива. Достатъчно е да

се потърси по „Е“ правило ключ в началото на изречение. Но и на друго място намерен, моделът допълва или пояснява определението. Затова в практическата реализация се извеждат и останалите случаи. Ето получени резултати от обработка на текстови извадки, върнати от ИТС по ключ „Алгоритъм“.



По подобен начин могат да бъдат реализирани и останалите четири отношения, описани в (Колковска, 2005): част–цяло, продукт–източник, функционални и каузативни.

Извежданите определения имат дедуктивна основа.

За по-задълбочен анализ на естественоезиков текст са необходими математически методи.

4. Възможности на литературно-честотния анализ. Сравнително лесна за реализация е статистическата честотна обработка на словоупотребите в различните функционални сфери на речта. Поради стремежа на издателите да използват ключовите думи по-често, документите в Интернет са удобно поле за изследване в тази посока.

Примерен алгоритъм за построяване на електронен честотен речник по дадена тема:

1. Всички текстови съдържания, върнати от ИТС, се събират в един документ.
2. Създава се масив на изреченията (символни низове, започващи с интервал, главна буква и завършващи с точка и интервал).
3. Създава се двумерен масив на свързаните (намиращи се в изречение) думи. В първа колонка се записват самите думи, а във втора – броят на участията.

4. Масивът се подрежда по азбучен ред.
5. Изключват се местоименията, предлозите и съюзите.
6. Думите се привеждат към еднаква коренна форма (обикновено най-кратката) с натрупване на участията (тежестите). По този начин се премахват множественото число, пълният и краткият член.

7. Масивът се преподрежда по брой участия.

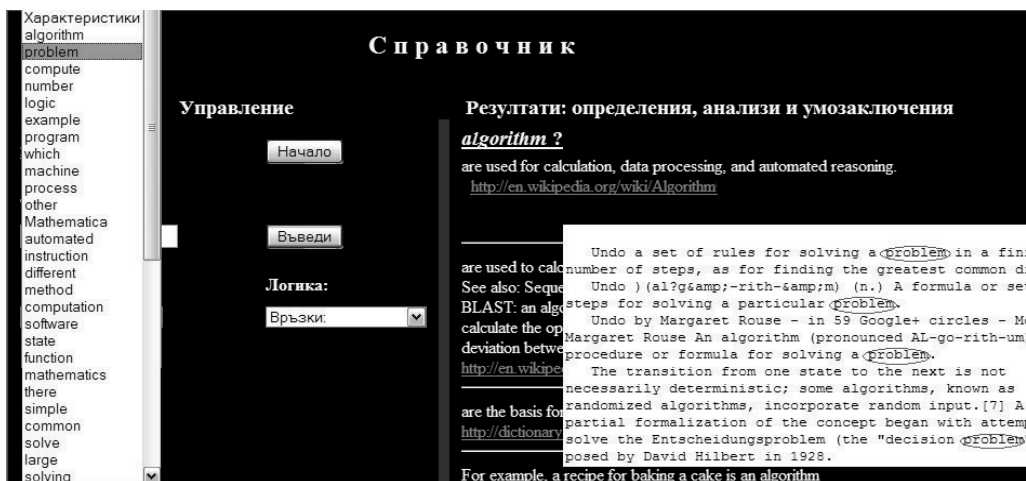
8. Първите 30 позиции се извеждат в списък „Участия“.

Ето попълнен списък на участията в ОТИС „Справочник“ (ключ „Алгоритъм“).
За по-голямо разнообразие сме запазили множественото число.



Забелязва се, че на първо място застава ключовото понятие, а след него се подреждат свойства и отношения, които го характеризират. Празните места (ако има такива) в списъка означават, че са елиминирани думи, които не са намерили потвърждение в свързан текст (не са по темата). Обективността на честотната оценка нараства в зависимост от обема и разнообразието на текстовете, върнати за анализ. Възможна е и обратната задача: извеждане на изреченията, в които се среща характеристиката, или получаване на предметен указател. Естествено е да се очаква, че обемът на посочената извадка ще намалява пропорционално на тежестта на избраната дума.

Системата се справя успешно и с английски език.



5. Класификатор на понятията. Ако изследваното понятие е ново за системата, се подготвя запис в „База знания“. Тъй като отношението род-вид има йерархична структура, под база знания тук ще разбираме „Класификатор на понятията“. Класификаторът е представен в MySQL таблица с полета: код, ниво, род, вид, характеристики и отношение.

Класификатор на понятията					
Код	Ниво	Род	Вид	Характеристики	Отношение
2501010001	6	успоредник	ромб	диагонал четири страни ъгли точка периметър лице триъгълници +++ математика >>> страни равни	1
2501010000	5	четириъгълник	успоредник	диагонал четири страни ъгли точка периметър лице +++ математика >>> страни успоредни	1
2501000000	4	геометрична фигура	четириъгълник	диагонал четири страни ъгли точка +++ математика >>> изъгънал	1

„Род“ съдържа по-общото или близко по смисъл намерено понятие. „Вид“ е въвежданото в йерархията понятие. Полето „характеристики“ се състои от две части: общи и специфични. За разграничаване се използват разделителни символи. В „отношение“ е кодирана връзката между базовото понятие и вида. Възприета е дървовидна организация на образците. Наредбата се задава посредством структурно-сериен код.

6. Извеждане на заключения по аналогия (синтез). За усвояване на словесното определение не е достатъчно само неговото запомняне. Необходимо е да се познава структурата на определението, да се овладеят правилата за построяване. Мисловните логически операции протичат по следната схема:

1. Понятието, което се дефинира, се приема за вид.

2. Подбира се по-общо по обем понятие (род).

3. Вид се подчинява на род.

4. Посочва се различното във вида.

В ОТИС „Справочник“ горната последователност е автоматизирана по следния алгоритъм:

1. Характеристиките от списък „Участия“ се разглеждат като множество.

2. Търси се пресичане на неговите елементи с поне пет от характеристиките на някой от образците в класификатора (критерият „пет и повече“ е експериментален и подлежи на уточняване).

3. Ако има такова, се подготвя ново знание:

3.1. Пресечените елементи се пренасят в новия запис по наредба, зададена в образаца.

3.2. Добавят се новите характеристики (липсващи в „Участия“).

3.3. Останалите характеристики от списъка се добавят като специфични.

3.4. Извежда се заключение по аналогия:

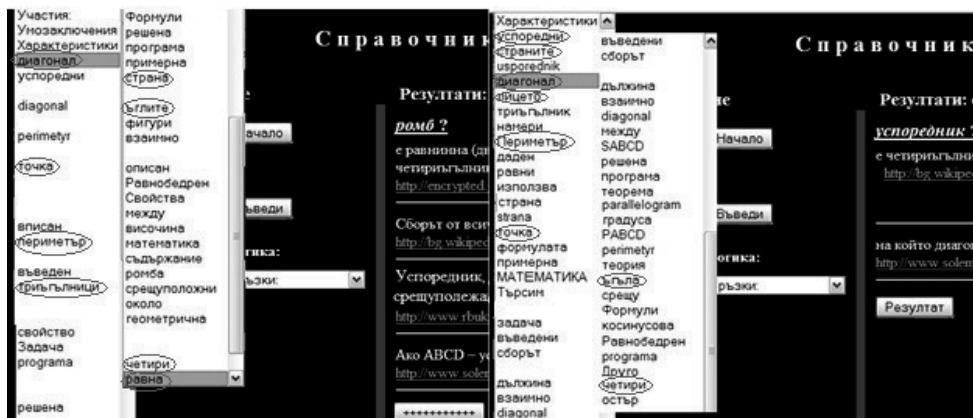
Y (род) притежава характеристики a, b, c, m

X (вид) притежава характеристики a, b, c, d

Следователно X е подобен на Y, като X вероятно притежава и m .

Новото знание се валидира от администратор. Ролята на администратора е да премахне случайните думи.

В демонстрационния пример ръчно бяха въведени записи за „геометрична фигура“ (предельно общо понятие) и „четириъгълник“. Записи за успоредник и ромб бяха създадени автоматично и бяха редактирани.



Справочник

Управление

За системата...

Начало

Тема:

ромб

Въведи

Анализ

Умозаключения ▾

Логика:

Връзки: ▾

Връзки:

ромб

успоредник

четириъгълник

геометрична_фигура

Резултати: определения, анализи и умозаключения

- лице и периметър Успоредник, в който всички страни са равни и всяка две срещуположни страни са успоредни ще наричаме ромб
<http://www.rbulvar.eu/romb.html>

Ако ABCD – успоредник и $AB = AD$ ABCD – ромб
<http://www.solembajr.com/Sam7Usp1.htm>

Резултат

Област математика.

успоредник притежава: диагонал четири страни ъгли точка периметър лице.

ромб притежава: успоредни диагонал триъгълник diagonal права равни периметър прави точка страни функции ъгли ъгли вписан периметър окръжност въведен триъгълници страна...

Следователно ромб е подобен на успоредник.

Пренесени характеристики от успоредник: лице.

Достоверността на извода зависи от:

- обективността на честотната оценка;
- броя на съвпадналите атрибути;
- местата на пресечените елементи в множеството на намерения образец

(колкото по-напред, толкова по-добре).

Йерархичната организация позволява проследяване на логическите връзки и проверка на останалите логически изводи по веригата (списък „Логика“).

Справочник

Управление

За системата...

Начало

Тема:

ромб

Въведи

Анализ

Умозаключения ▾

Логика:

успоредник ▾

Връзки:

ромб

успоредник

четириъгълник

геометрична_фигура

Резултати: определения, анализи и умозаключения

четириъгълник с четири равни страни.
<http://encrypted.google.com/search?q=romb>

Сборът от всички ъгли на ромба е 360 градуса
<http://bg.wikipedia.org/wiki/Ромб>

- лице и периметър Успоредник, в който всички страни са равни и всяка две срещуположни страни са успоредни ще наричаме ромб
<http://www.rbulvar.eu/romb.html>

Ако ABCD – успоредник и $AB = AD$ ABCD – ромб
<http://www.solembajr.com/Sam7Usp1.htm>

Резултат

Област математика.

успоредник е четириъгълник (вид - род).

Общи характеристики:
диагонал четири страни ъгли точка периметър лице

Специфични характеристики: страни_успоредни

7. Заключение. ОТИС „Справочник” е в експериментален период и се ползва индивидуално от учители и като помощно техническо средство в часовете по Информатика и Информационни технологии в СОУ „П. Волов” – Шумен. В случаите, когато се налага въвеждане на ново понятие, тя се оказва верен помощник на учителя, поддържа непрестанен интерес, повишава мотивацията и в крайна сметка води до по-качествено обучение. Налице е тясно интегриране между двата предмета.

При опитната експлоатация излязоха наяве и някои особености. Например търсене по „релеф“ смесва термини от географията с такива от фитнес, бодибилдинг, културизъм. Но ограниченото по обем „релеф география“ отделя правилно съществените думи. Естествено, при формиране на заявката тук важат правилата, валидни и за ИТС.

В дългосрочен план, освен реализиране на останалите КСО, се предвижда добавяне на средства за обработка и на чуждоезиков текст.

БЕЛЕЖКИ

Колковска, С. (2005). Модели на концептуално-семантичните отношения в специален (химически) текст с оглед на автоматичното им разпознаване. http://liternet.bg/publish16/s_kolkovska/modeli/05.htm

ЛИТЕРАТУРА

- Иванова, Т. & Момчев, И. (2008). Персонализирана уеб търсеща система, базирана на семантични и многоагентни технологии. *Автоматика и Информатика*, 3, 29–33.
- Попов, Э. В. (Ред). (1990). *Искусственный интеллект – справочник в трех книгах. Книга Радио и связь*. Москва.
- Мельников, В. Н. (1989). *Логические задачи*. Киев – Одесса: Высшая школа.

✉ **Ивайло Пенев**

асистент

Департамент по компютърни науки и технологии, ТУ, Варна

9000 Варна

E-mail: ivailo.penev@tu-varna.bg

✉ **Пламен Пенев**

учител

училище „Панайот Волов”

ул. Съединение 100, 9700 Шумен

E-mail: plampenev@abv.bg

EXTRACTION OF OBJECTIVE KNOWLEDGE FROM INTERNET

Abstract. The article offers an approach and utilities for extraction of important information from Internet on a subject interesting to us. A demonstrative generalizing search system „Handbook“, containing elements of semantic analysis, is described. A simple technique for obtaining deductive definitions is realized. Frequency word book and an index are constructed. Storing the keywords into knowledge base allows production of logic conclusions by analogy.

✉ **Ivaylo Penev**

Assistant

Department of Computer Sciences and Technologies, Technical University of Varna
9000 Varna, Bulgaria

e-mail: ivailo.penev@tu-varna.bg

✉ **Plamen Penev**

Teacher

„Panayot Volov“ School
100, Soedinenie Street, 9700 Shumen, Bulgaria

e-mail: plampenev@abv.bg