

PROSTŘEDKY PRO ZPŘÍSTUPNĚNÍ A VYHLEDÁVÁNÍ TEXTOVÝCH INFORMACÍ

Václav Snášel, Jiří Dvorský ^{a)}
Petr Šaloun, Daniela Ďuráková ^{b)}

^{a)} Univerzita Palackého, Tomkova 40, 779 00 Olomouc

^{b)} VŠB – Technická univerzita, 17. listopadu 15, 708 33 Ostrava

Abstrakt

Současná doba je charakterizována prudkým nárůstem potřeby a vyhodnocení informací. Vývoj a zdokonalování dokumentografických informačních systémů je stále předmětem výzkumu. Mezi základní problémy patří řešení vyhledávacího problému, to znamená k uživatelskému požadavku přiřadit odpovídající dokumenty. V tomto článku jsou diskutovány postupy při realizaci informačního systému Agrokrom, v němž využíváme spojení fulltextových, hypertextových a databázových technologií.

1. Úvod a motivace

V současné době neustále stoupá potřeba zpracování velkého množství informací – novinových článků, odborné literatury, korespondence, agenturních zpráv, vyhlášek, zákonů, příspěvků z konferencí na počítačových sítích atd., které jsou přístupné v textovém tvaru (dále budeme používat pojem text). Jedná se většinou o stovky nebo tisíce stran obsahujících nejrůznější informace, z nichž však je pro konkrétního čtenáře v mnoha případech zajímavý pouze nepatrný zlomek. Ostatní stránky textu nejsou pro něj potřebné, protože obsahují informace, které již zná a nebo nepatří do oblasti jeho zájmu. Na první pohled je tedy všechno jednoduché. Během několika málo minut se čtenář seznámí s tím, co ho zajímá a zbytkem se vůbec nezabývá. Problém ale spočívá v tom, že většinou neví, kde přesně se pro něj zajímavé informace nacházejí nebo dokonce ani neví, zda se vůbec žádané informace v dostupných textech nacházejí.

Potřeba počítačového zpracování textů je tedy velmi aktuální. To lze dokumentovat i tím že podle statistik zabere vyhledávání informací v takovýchto dokumentech určeným pracovníkům až dvě hodiny pracovního času denně.

V příspěvku popisujeme spolupráci pracovníků katedry informatiky FEI VŠB—TU Ostrava, katedry matematické informatiky PřF UP Olomouc a Zemědělského výzkumného ústavu Kroměříž, s.r.o, dále zkráceně VUKROM. Spojením odborného zemědělského subjektu a informatických univerzitních pracovišť vznikl tým optimálně využívající jednotlivé specializace. Komerční subjekt spolupráce je zárukou praktické realizace a skutečného nasazení a provozu vytvořených produktů i know how.

Jedním z výsledků spolupráce je aplikace Agrokrom, která podporuje rozhodování uživatele bohatými znalostními databázemi v oboru a současně nabízí přímý přístup k vybraným textovým a obrazovým informacím, které jsou kontextově spojeny s aktuálně řešenou odbornou problematikou. Bližší informace o vlastnostech aplikace Agrokrom již byly publikovány v časopise Obilnářské listy a jsou dostupné na internetu, viz [8].

Tento příspěvek stručně popisuje databázovou část aplikace, hlavně se však zaměřuje na část umožňující přípravu, zpracování, publikování, prezentaci a vyhledávání ve vědeckotechnických a zemědělských informacích. V příspěvku rovněž popisujeme vytvořené techniky a vlastní programové nástroje, které uvedené činnosti podporují.

2. Dostupné informace

Odborný i kvantitativní rozsah informací dostupných a souvisejících s aplikací Agrokrom nám umožňuje provést jejich zjednodušený rozklad na níže uvedené oblasti.

- Soubor textů popisujících vybrané plevely, choroby a škůdce v rostlinné výrobě doplněný řadou fotografií a perokreseb, etikety mnoha registrovaných pesticidů a popisy vlastností většiny odrůd nejen obilovin, ale i dalších plodin.
- Soubor metodických příruček o pěstování ozimých a jarních obilovin, brambor, cukrovky a dále soubor rámcových metodik pěstování řady plodin.
- Soubor vybraných článků publikovaných v časopise *Obilnářské listy* a dále soubor článků a textů o speciálních problémech v rostlinné výrobě.
- Soubor textů v oblasti ekonomiky. Soubor je svým způsobem ojedinělý, neboť mimo jiné prezentuje výkladový slovník pojmů v ekonomice a je publikován v česko-německé verzi, stejně tak jako katalog ekonomických pojmů pro poradce. Součástí tohoto celku je i manuál pro poradce objasňující metodiku a principy postupů v oboru *Ekonomická optimalizace hospodaření zemědělských podniků*. Konečně celek obsahuje i texty o účetním a manažerském pojetí nákladů.
- Soubor vybraných citací vědeckých, výzkumných a odborných článků podporovaný výkonným prohledávačem podle zvolených kritérií významně posiluje informační složku celého systému.
- Databáze informací pro podporu rozhodování zemědělských subjektů při jednotlivých pracovních postupech spojených s pěstováním kulturních plodin na obhospodařovaných pozemcích.

Po nezbytném uvedení do problematiky a rozsahu informací již další text zaměříme informaticky.

3. Volba technologií

Úvodem této části poznamenejme, že popisované rozhodování časově spadá do poloviny roku 1997. Přesto se domníváme, že zvolené technologie odpovídají i současným trendům a možnostem.

Spojení textu, fotografií, obrázků a odkazů (hyperlinků) vyžadovalo použití univerzálního formátu pro prezentaci v podobě, která je shodná se vzhledem na straně tvůrce. Tím byl jednoznačně vyloučen formát HTML, často deformující vzhled podle nastavení prohlížeče. Dále byl vyloučen formát DOC populárního Wordu. Jeho prohlížeče od firmy Microsoft jsou sice použitelné bezplatně, bohužel rovněž ony nezajistí identický vzhled zobrazení s původním dokumentem. Volba padla na formát PDF (Portable Document Format) firmy Adobe, který přesně splňuje požadované vlastnosti. Tento formát navíc spojí všechny informace do jediného celku, čímž odpadá problém správy HTML souborů a do nich začleňovaných obrázků. PDF formát je navíc vhodný pro použití v prostředí inter/intranetu. PDF dokonce nabízí možnosti ochrany dokumentu, což je významné. Neboť prezentované texty i fotografie byly pořízeny za značné částky. PDF prezentuje informace kvalitně a přitom neposkytuje jejich originální podobu.

O správnosti této volby svědčí obliba publikování starších čísel časopisů právě v PDF formátu. Příkladem necht' jsou periodika nakladatelství Computer Press. V našem případě jde o *Obilnářské listy*, jejichž starší čísla jsou umístěna na [8], ale i výzkumné zprávy a odborné články.

Pro databázovou část aplikace bylo zvoleno prostředí WinBase 602 firmy Software602. Rozhodujícími aspekty byly architektura klient/server, (ne)náročnost na hardware, možnost vývoje aplikace týmem programátorů v síťové verzi, technická podpora výrobce vývojového prostředí a v neposlední řadě to byla i poměrně příznivá cenová politika firmy Software602.

Po ročním vývoji vznikla první verze komplexního informačního a expertního systému pro zemědělce Agrokrom, který zajišťoval nejen databázovou evidenci údajů pro rostlinnou výrobu, ale umožňoval i přímý přístup uživatelů k podrobným informacím (popsaných v předešlé kapitole) pomocí kontextových odkazů. Uspořádání informací zachovává odborné členění. Přitom je využito vlastností souborového systému, kdy jednotlivé podadresáře (složky) odpovídají hierarchii členění. Nezávisle na tomto přístupu byla zvolena progresivní technologie fulltextového vyhledávání. Ta umožňuje snadno nalézt všechny dokumenty, obsahující jakýkoliv výskyt hledaného řetězce s případnými dalšími podmínkami. Tedy například slovo „mrkev“ jak v kontextu plodin, ale i hnojení, ochrany před škůdci a dalších, napohled ne zcela zřejmých souvislostech. Podrobněji se fulltextové problematice věnujeme dále v tomto příspěvku.

4. Uživatelské prostředí

Uživatelské prostředí respektuje návyky a možnosti cílové skupiny uživatelů. Expertní a informační systém Agrokrom tedy pracuje v prostředí českých Windows NT 4.0, Windows 95/98 a Windows 2000. Databázově je postaven nad WinBase602 a vyžaduje přítomnost Adobe Acrobat Readeru.

5. Databázová aplikace Agrokrom

Aplikace Agrokrom představuje klasickou relační databázi, která odborníkům pro rostlinnou výrobu zajišťuje nejen evidenci potřebných údajů, ale také zpřístupňuje podrobné informace nutné při ekonomickém hodnocení možných variant různých pěstebních plánů. Tvorba systému je popsána též v [9].

Vzhledem k rozsáhlosti zpracovávaných údajů byl od počátku vývoj aplikace rozdělen do několika modulů, které obsahovaly data související s jednotlivými pracovními operacemi v rostlinné výrobě. Modulární členění představovalo také výhody při postupném zpracování analýzy a rozdělení implementace systému mezi tým programátorů.

V prostředí WinBase602 verze 5.1 se však volba modulárního členění při sestavování modulů do jedné aplikace ukázala jako omezující faktor. Každý z modulů byl chápán jako jedna samostatná databáze. Přitom charakter určitého množství dat byl společný všem modulům, což vyžadovalo propojení mezi nimi. Testováním přístupu přes ODBC se zjistilo, že dochází k velkým časovým prodlevám při přenosu dat mezi moduly. Bylo použito rozhraní, které zajistilo přístup ke všem modulům a obvyklé ovládací prvky.

Pro jednotné uživatelské rozhraní byl proto použit standardní prohlížeč Microsoft Internet Explorer 4.1. Rozhraní bylo vytvořeno formou HTML stránek s vloženými ActiveX propojeními na hlavní pohledy jednotlivých modulů., ve kterých byly připraveny operace pro zpracování dat.

Aktuální verze Agrokromu 2.0 již přítomnost MSIE nevyžaduje, neboť zlepšení vývojových prostředků ze strany WinBase602 verze 6.0 umožnilo zpracovat moduly jako

jedinou databázovou aplikaci. Uživatelské rozhraní je vytvořeno formou nabídkového menu, které umožňuje přístup ke kterékoli části databáze.

Obě verze aplikace Agrokrom slouží k vyhledávání informací klasickými databázovými prostředky a podporují přímý přístup k souborům podrobných textových informací. Vybrané prvky v databázi obsahují jako jednu z popisujících položek odkaz na odpovídající soubor formátu PDF. K zobrazení těchto informací není nutné opustit prostředí aplikace Agrokromu, zobrazení je zajištěno pomocí nainstalovaného Adobe Acrobat Readeru.

6. Vývojové prostředí

Vývojové prostředí musí do jisté míry vycházet z prostředí uživatelského. Oproti prostředí uživatelskému však musí být robustnější co se stability týče a musí spolupracovat s vývojovými nástroji. Tento přirozený požadavek je bohužel v rozporu s realitou prostředí a aplikací pro Windows NT a 95/98. Například zatímco pro pořizování a zpracování datově rozsáhlých fotografií je výhodný skener připojený přes SCSI rozhraní, dostatek operační paměti, rychlý disk a procesor, Adobe Photoshop a WinNT, je pro správnou tvorbu PDF nutný Adobe Acrobat (Distiler) spolu s ATM a českými postscriptovými fonty, ovšem v prostředí Win 95/98. S úspěchem jsme využili Microsoft Word s makrem PDFMaker.

Vývojové prostředí musí ovšem vycházet i z formátu podkladových informací. Je totiž jasné, že tvůrci aplikace nejsou současně autory fotografií a odborných textů. Ty jsou získány i od řady spolupracovníků. Pro zajištění jednotného vzhledu skupin dokumentů je nutné spolupracovníky proškolit nejen v použití zvoleného editoru Microsoft Word, ale i v základech počítačové sazby a typografie. Současně bylo nutno připravit používané styly a makra. Výběr fontů a příprava stylu byly podřízeny záměru poskytnout perfektní a věrné zobrazení na uživatelské obrazovce při zachování estetické úrovně vytištěného dokumentu.

7. Použití fulltextové technologie

Důležitou součástí projektu je realizace textové databáze. Tato databáze byla realizována pomocí fulltextového jádra využívaného v projektu Amphora viz [6]. Na úvod stručně přiblížíme problematiku textových databází. Další podrobnosti lze nalézt v [4, 5].

8. Dokumentografický informační systém, dokument

Je zřejmé, že nás nebude zajímat pouze část textu vytržená z kontextu, ale ucelený text obsahující hledanou informaci. Takovýto text budeme dále nazývat dokumentem.

Tento pojem bude základem pro všechny činnosti související se zpracováním textových informací.

Na dokument se můžeme dívat ze dvou pohledů:

- Fyzický pohled určuje způsob uložení dokumentu na médiu.
- Logický pohled je dán informacemi, které jsou v dokumentu obsaženy.

Logický dokument nemusí odpovídat jednomu fyzickému dokumentu. Například kniha jako jeden logický celek může být rozdělena do více fyzických souborů.

9. Dokumentografické informační systémy

Dokumentografické informační systémy jsou třída programových nástrojů, určených pro zpracovávání, úschovu a výběr dokumentů.

Základním rozdílem mezi dokumentografickými a faktografickými informačními systémy je strukturovanost vkládaných dat. Faktografické systémy pracují s daty, majícími pevnou, předem danou strukturou, ve které každá položka má předem daný význam. Příkladem faktografických systémů jsou v současné době rozšířené relační databázové systémy.

Dokumentografické informační systémy (DIS) naproti tomu pracují s daty, která jsou ve své podstatě strukturována jen málo, nebo vůbec ne. Základním prvkem dat v těchto systémech je text v přirozeném jazyce. Může se jednat o zákony, o knihy, o časopisy, o úřední spisy, o diplomové práce stejně jako o výzkumné zprávy, shrnující výsledky vědeckých pokusů. Například Institut pro standardizaci a technologii USA má k dispozici více než dva milióny textových dokumentů viz [2]. Po vzoru faktografických informačních systémů budeme na uložená textová data nahlížet jako na (textovou) databázi.

Nízká strukturovanost uchovávaných dat a použití přirozeného jazyka přináší nutnost vyřešit při tvorbě DIS i řadu jazykových problémů.

10. Obecné schéma DIS

Vstupní texty jsou po předzpracování zahrnuty do textové databáze, což je struktura obsahující informace o textech v podobě vhodné k aplikaci vyhledávacích technik. Uživatel klade dotaz, který je následně podroben analýze syntaktické a sémantické, pak je transformován a stává se vstupem pro algoritmus vyhodnocení dotazu. Výstupem tohoto algoritmu je nějaká množina záznamů, kterou je třeba dále zpracovat (seřazení, ohodnocení, dohledání vzorků atd. ...). Výsledné dokumenty jsou pak předloženy uživateli, který se rozhodne, zda se s dosavadní odpovědí spokojí, či zda bude pokračovat dalším, vylepšeným, dotazem, pomocí zpětné vazby. Vstupem algoritmu pro konstrukci zpětnovazebního dotazu je výsledek předchozího vyhledávání spolu s údajem uživatele, které vybrané dokumenty považuje za relevantní.

DIS se skládá z několika spolupracujících komponent viz [5]. Ne v každé implementaci se vyskytují všechny zde popsané části. Zdokonalování všech těchto komponent je stále ještě předmětem výzkumů. Kvalita jednotlivých prvků, použitých v konkrétním systému potom určuje výslednou kvalitu, tedy míru uspokojení jeho uživatele či uživatelů.

Jednou z komponent DIS může být vstupní textový filtr, který provádí lingvistickou analýzu čteného textu a převádí přečtené lexikální jednotky textu (slova) na základní tvar. Nazývá se často *lematizátor*.

V existujících jazycích (český jazyk je v tomto směru velmi bohatý) má většina slov mnoho tvarů, lišících se podle rodu, pádu, jednotného či množného čísla. Jindy může mít jedno slovo několik odlišných významů (např. let = rok, léta).

Z našich testů vyplývá, že před indexováním nemá význam upravovat vstupní text filtrací. To proto, že filtrací dochází ke ztrátě informace a u velmi rozsáhlých textových databázích nedojde ani k očekávané úspoře rozsahu indexů. Podstatně lepší výsledky lze dosáhnout kompresí textového dokumentu viz [3] a kompresí indexů.

Další komponentou v klasických DIS je indexační jednotka. Tato komponenta má za úkol obohatit ukládané texty o doplňující informace, které umožní efektivní vyhledávání. V této fázi zpracování se ke každému textovému dokumentu doplní jeho počítačová reprezentace, která se nazývá záznam dokumentu.

Záznam obsahuje formální popis dokumentu, skládající se z hodnot vhodně specifikovaných atributů (položek), a z množiny termů, které ve stručné podobě vystihují obsah plného znění dokumentu. Vzhledem k nejednotné terminologii budeme pod termem chápat jistý vzorek textu (výraz), který může být víceslovný nebo také jednoslovný. Jednoslovným (ale

mnohdy i víceslovným) termům se také někdy říká klíčová slova, místo o termech se také hovoří o deskriptorech.

Nalezení vhodné množiny termů je obecně velmi náročná úloha, která v mezním případě vyžaduje porozumění sémantickému významu textu. Termy, vybrané během indexace musí dostatečně přesně reprezentovat obsah dokumentu a také dát do souvislosti dokumenty, týkající se podobného tématu. Přitom obecné termy, vyskytující se ve všech, resp. skoro ve všech dokumentech, nemají pro účel vyhledávání téměř žádný význam. Malý význam mají také ty termy, které se vyskytují v příliš malém počtu dokumentů. Tento problém se řeší pomocí *StopListu*. StopList je seznam slov, která se při indexování a dalším zpracování textu ignorují. Konstrukce stoplistu je uvedena v [2, 3].

Proces přiřazení množiny termů dokumentu – indexace dokumentu – se proto v mnoha systémech provádí buď ručně, nebo poloautomaticky. V prvním případě provede specialista v daném oboru – indexátor – sám výběr nejvhodnějších termů, v druhém případě systém poskytuje možnost upravit množinu termů, vytvořenou systémem na základě analýzy plného textu.

V [1] je popsán pokus, kdy osm indexátorů mělo pro popis dokumentu vybírat ze čtrnácti deskriptorů. Ukázalo se, že žádný deskriptor nebyl vybrán všemi, a žádní dva indexátoři nepoužili stejnou množinu deskriptorů.

Jak již bylo uvedeno v úvodu příspěvku, jsou zdrojové dokumenty pořizovány ve Wordu a cílové dokumenty jsou distribuovány ve formátu PDF. Ukázalo se proto jako velmi výhodné svázat modul pro indexování i s patřičnými konverzemi. Vlastní indexování probíhá podle volby ve třech fázích:

- převod dokumentu ve formátu Word do textu;
 - převod dokumentu ve formátu Word do formátu PDF a nahrazení odkazů z wordovských dokumentů na dokumenty PDF;
 - vlastní indexování PDF dokumentů.
- Kolekce indexovaných dokumentů obsahuje přibližně 2000 dokumentů.

11. Metody vyhledávání

Při vyhledávání v dokumentech se vychází ze slov, která charakterizují jeho obsah. Tato slova mohou být stanovena uměle, například autorem dokumentu a mají pak podobu klíčových slov. To však přináší několik problémů. Například nemusí být jasné, kolik slov dostatečně charakterizuje obsah dokumentu. Volba klíčových slov je navíc značně závislá na pohledu jejich tvůrce.

Přesnější je tedy, když pracujeme s celým dokumentem a uvažujeme všechna slova, která jsou v něm uvedena. Ani to samozřejmě nezaručuje, zaznamenání všech informací obsažených v dokumentu. Jeden logický pojem můžeme vyjádřit pomocí různých slov. Hledáme-li například dokumenty obsahující informaci o kopané, tak nás budou zajímat i dokumenty, ve kterých se slovo kopaná vůbec nevyskytuje, například dokument obsahující slovo fotbal. Je celkem pravděpodobné, že nás mohou zaujmout i dokumenty obsahující slova Sparta, Slavie, FIFA, penalta, ...Všetchna tato slova totiž mohou (ale také nemusí) souviset s naším dotazem.

V DIS lze pak formulovat základní vyhledávací problém:
Nalézt k uživatelskému požadavku – dotazu relevantní dokumenty.

Mezi problémy, souvisejícími s vyhledávacím problémem patří zejména:

- jak určit, co je relevantní a co ne,
- jak zajistit efektivnost zpracování,
- jak zajistit uspořádání výstupů podle relevance.

Je zřejmé, že řešení vyhledávacího problému vyžaduje další komponentu DIS – vyhledávací stroj. Tato komponenta využívá indexů a vybírá z textové databáze dokumenty, které vyhovují dotazu, zadanému uživatelem.

Vyhodnocení dotazu spočívá většinou v porovnání termů uvedených uživatelem v dotazu s popisy, které specifikují jednotlivé dokumenty.

Indexace ovšem nemusí být vyjádřena pomocí termů. Obsah dokumentu lze vyjádřit i jinak, například jistým zakódováním textu do podstatně kratšího řetězce znaků – signatury. Jinou možností může být ocenění termů popisujících dokument číslly (váhami) vyjadřujícími důležitost termu v dokumentu. Využití lingvistiky může znamenat např. využití jistých relací mezi termy. Takové relace tvoří další pomocné struktury dat (tezaury), které mohou zaručit výběr takových dokumentů, jejichž termy jsou jiné, než ty zadané v dotazu, a přesto je výsledek relevantní.

Zobecníme-li tuto úvahu, potřebujeme model textové databáze. Bude to soubor pojmů a nástrojů umožňujících popsat textovou databázi a formulovat základní vyhledávací algoritmy umožňující řešit vyhledávací problém.

Ze softwarově inženýrského hlediska nesmíme zapomenout na komponentu uživatelské rozhraní. Ta komunikuje s uživatelem a nabízí mu možnost pokládat dotazy informačnímu systému za pomoci dotazovacího jazyka. Zatímco v oblasti faktografických informačních systémů existují standardy pro komunikaci (nejznámějším dotazovacím jazykem je SQL), kterými se výrobci řídí, standardizace dotazovacích jazyků v textových databázích je teprve v začátcích a každý produkt obsahuje vlastní způsob formulace dotazů.

12. Booleovský model

Ve většině případů jsou dotazy formulovány pomocí přesně definovaného formálního jazyka (založeného většinou na Booleově algebře). Booleovská metoda dotazování je nejstarší a také nejrozšířenější způsob formulace dotazu uživatelem. V tomto modelu je každý dokument spojen s množinou termů, kterými je charakterizován a dotaz je booleovský výraz (složen z termů, logických operací AND, OR, NOT a uzávorkování). Z databáze jsou vybrány ty dokumenty, které obsahují hledané termy v kombinaci určené dotazem. Vyhodnocování těchto dotazů je založeno na vyhledávání v tzv. invertovaných seznamech (pro každý term máme uspořádaný seznam dokumentů, ve kterých se vyskytuje).

Uvedeme popis operací pro booleovský model.

X AND Y Výběr dokumentů, obsahujících jak term X, tak term Y.

X OR Y Výběr dokumentů, obsahujících buď term X, nebo term Y, nebo oba termy současně.

X XOR Y Výběr dokumentů, obsahujících buď term X, nebo term Y, ale ne oba současně.

X NOT Y Výběr dokumentů, obsahujících term X, ale ne term Y.

X ADJ Y (adjacent) - Výběr dokumentů, ve kterých se vyskytuje term X následovaný termem Y

X WORDS(n) Y Výběr dokumentů, ve kterých se vyskytuje term X následovaný termem Y nejdále ve vzdálenosti n slov.

Tento model se velice snadno implementuje a je velice efektivní z hlediska časové náročnosti na vyhodnocení dotazu.

13. Zadávání fulltextového dotazu

Na základě zkušeností s realizací textových databází, viz [7], jsme se rozhodli pro dvě varianty dotazu.

- Jednoduchý dotaz.
- Přímý dotaz.

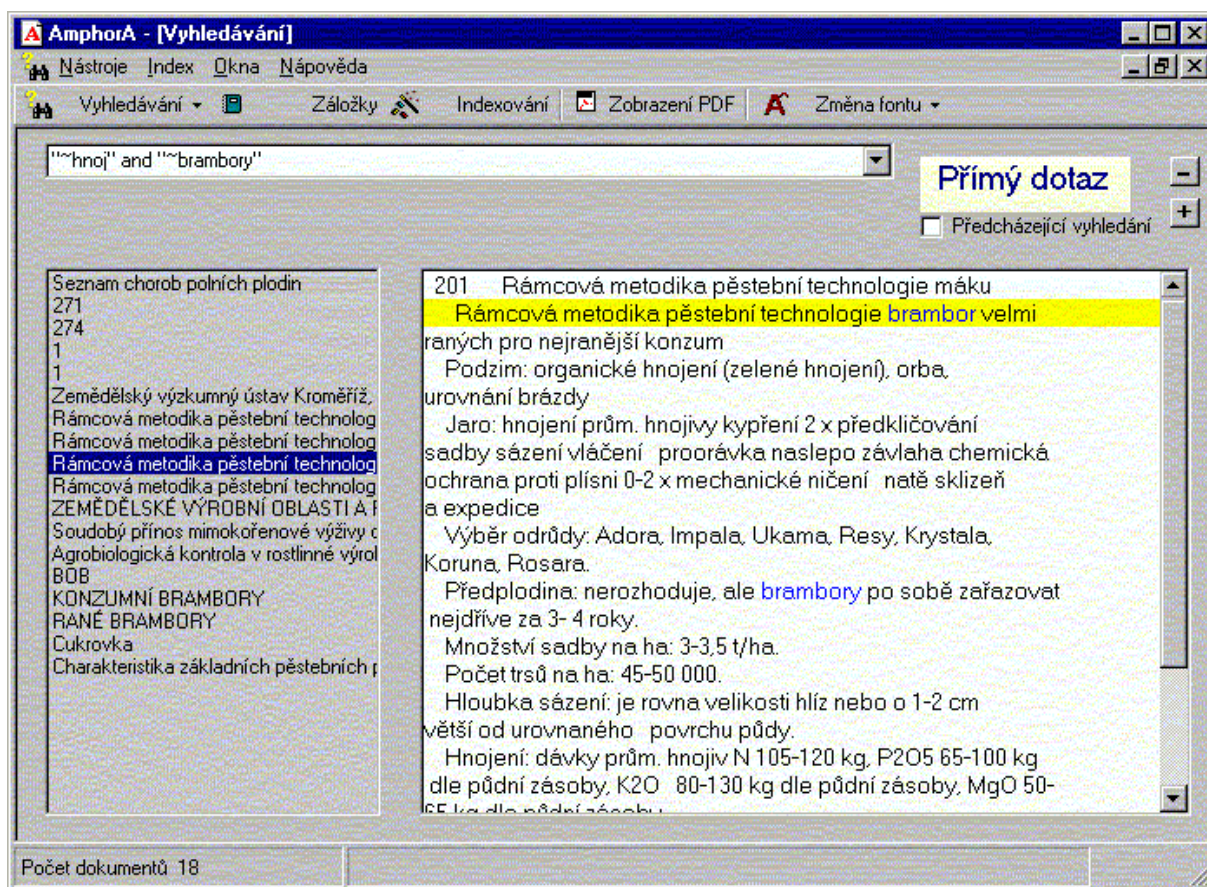
Přímý dotaz je booleovský výraz, který kombinuje jednotlivá slova, regulární výrazy nebo lemata. Lema se zadává pomocí znaku '~' viz obrázek 1.

14. Závěr

Popsaná spolupráce využívá grantové podpory ministerstva zemědělství EP 0960006075 *Informační systémy pro rostlinnou výrobu a expertní systémy pro hodnocení podnikatelských záměrů a finanční prognózu zemědělských podniků* a ministerstva školství Infra2, projekt LB98227 *Zpřístupnění distribuovaných informací v rámci INTRANETu a INTERNETu v oblasti zemědělských a technických věd*. Hlavním řešitelem je Zemědělský výzkumný ústav Kroměříž, s.r.o., spoluřešiteli či spolupracovníky jsou i autoři tohoto příspěvku a jejich pracoviště (katedry).

Budoucnost spolupráce vidíme ve dvou oblastech:

- integrace znalostní databáze a textové databáze pomocí XML,
- rozšiřování systému o multimediální prvky a prostorová data GIS.



Obrázek 1 Přímý dotaz

Literatura

1. D.C.Blair. Indeterminacy in the Subject Access to Documents., Information Processing & Management, Vol.22, No.2, 1986.
2. W.F.Frakes, R.B.Yares Ed. Information Retrieval, Data Structures & Algorithms Prentice Hall 1992.
3. R.Baeza-Yates, B. Riberio-Neto. Modern Information Retrieval. Addison Wesley, 1999.
4. B.Melichar. Textové informační systémy. Skriptum ČVUT, Praha 1994.
5. J.Pokorný, V.Snášel, D.Húsek. Dokumentografické informační systémy. Karolinum, Skriptum MFF UK Praha, 1998, ISBN 80-7184-764-X. pp 158.
6. V.Snášel, V.Sklenář, R.Nováková. Úplné texty a informační systémy. DATASEM 94.
7. V.Snášel, J.Urbánek. Zkušenosti s využitím a realizací fulltextového systému. Práce s externími datovými zdroji. Ostrava 1996, ISBN 80-902142-3-1, 21-26.
8. www.vukrom.cz informace o aplikaci Agrokrom a aktuálním stavu projektu Infra2.
9. D.Řuráková, P.Šaloun. Agrokrom, zemědělský informační systém. DATASEM 99.