



ZPŮSOBY HODNOCENÍ RELEVANCE U POKROČILÝCH VYHLEDÁVACÍCH SYSTÉMŮ



Pavel Kocourek
KMA INCAD



Terminologie na začátek

- Řeč bude o kategorii „Enterprise search“
(pokročilé ES – využití významového a podobnostního vyhledávání)
- RR – hodnocení relevance dokumentů k
dotazu (graficky nebo číselně prezentováno)

*Algoritmy hodnocení relevance rozlišují vyhledávací řešení -
details jsou mnohdy utajované, obecné způsoby hodnocení
však ne ...*



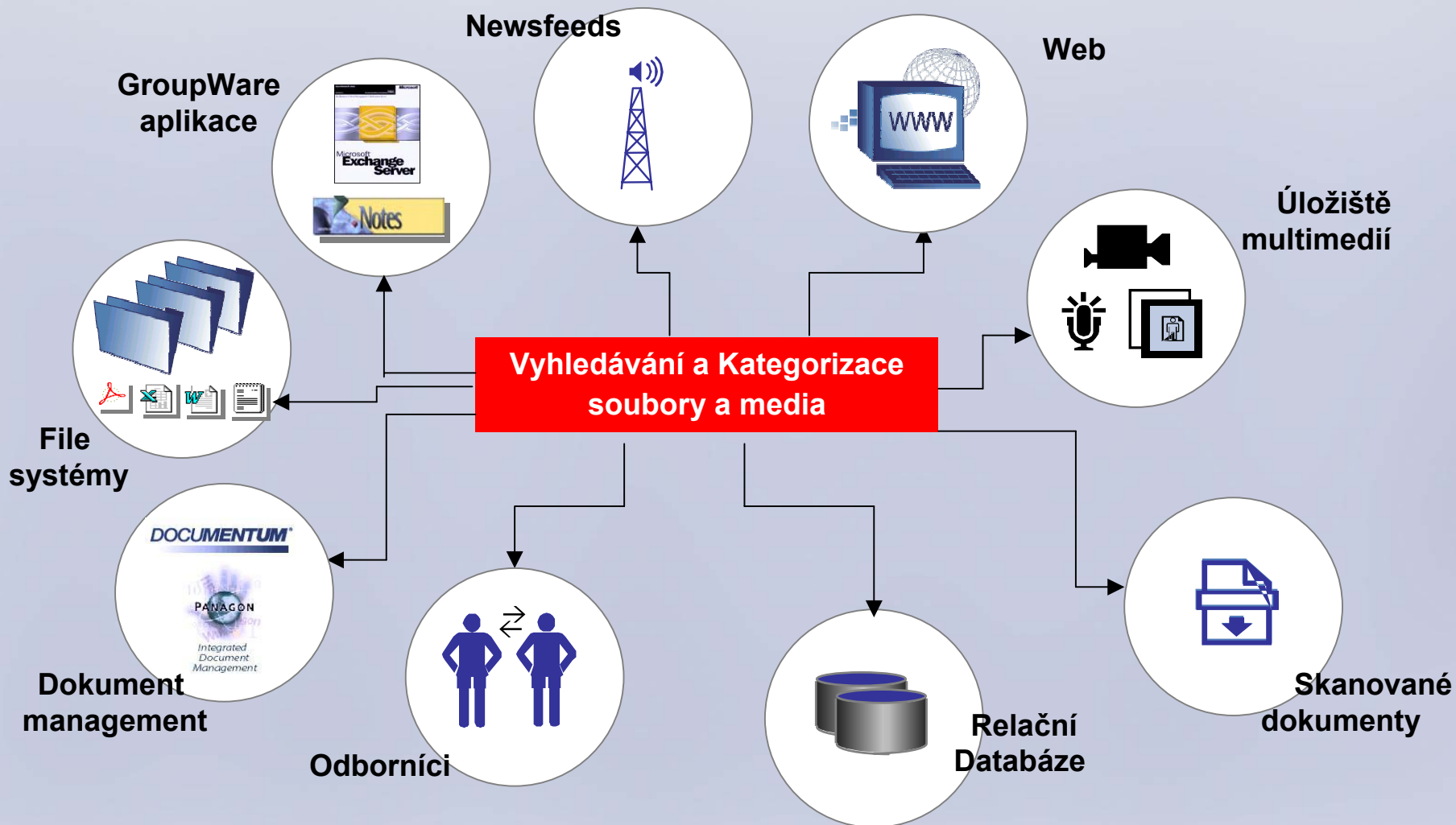
Situace mimo aulu ☺





Proč potřebuji vyhledávání s RR ...

Nevím kde jsou informace uloženy ...



Nevím v jakém jsou formátu ...

Microsoft Access * Adobe Acrobat * AMI Ami Pro Ami Professional Write Plus * BDR Microsoft Office Binder * BMP Windows Bitmap OS/2 Bitmap OS/2 Warp Bitmap Windows Cursor Windows Icon Corel Draw 2.0 Corel Draw 3.0 Corel Draw 4.0 Corel Draw 5.0 * CGM Computer Graphics Metafile * DBS DBase III DBase IV DBase V * DEZ DataEase 4.x * DIF Navy DIF * DRW Micrografx drawing products * DX DEC DX 3.0 and below DEC DX 3.1 DEC DX 4.x * DXF AutoCad Interchange ASCII AutoCad Interchange Binary * Documentum * EN4 Enable word processor 4.x * ENS Enable Spreadsheet * ENW Enable word processor 3.0 * EXE2 DOS Executable Windows Executable or DLL * FAX CCITT Group 3 Fax * FCD First Choice DB * FCS First Choice SS * FFT IBM DCA/FFT * FLW Freelance 1.0 & 2.0 for Windows Freelance 96 for Windows 95 Freelance 1.0 & 2.0 for OS/2 * FWK Framework III * FileNet Panagon * GIF Compuserve GIF * HGS Harvard Graphics DOS 3.0 Chart Harvard Graphics DOS 2.0 Chart Harvard Graphics DOS 3.0 Present. * HTML Internet HyperText Markup Language * IMG GEM Image * IWP Wang IWP * INFORMIX * JW JustWrite 1.0 JustWrite 2.0 Q&A Write 3 * LEG Legacy Wordstar for Windows * LZH LZH Compress LZA Self Extracting Compress * M11 Mass 11 * MANU Lotus Manuscript 1.0 Lotus Manuscript 2.0 * Lotus Notes/Domino *MCW MacWrite II * MM MultiMate 3.6 MultiMate Advantage 2 * MM4 MultiMate 4.0 * MMFN MultiMate Note * MP Multiplan 4 * MSW Microsoft Word 4.x Microsoft Word 5.x Microsoft Word 6.x Windows Write * MWKD Mac Works 2.0 Database * MWKS Mac Works 2.0 Spreadsheet * MWP2 Mac WordPerfect 2.0 Mac WordPerfect 3.0 * MWPF Mac WordPerfect 1.x * MWRK Mac Works 2.0 WP * OMF - OS/2 Only OS/2 Metafile * OW OfficeWriter * ORACLE * PCL PC File 5.0 Doc * PCX Paintbrush DCX (multipage PCX) * PDX Paradox 2 & 3 Paradox 3.5 Paradox 4 Paradox for Windows * PFS PFS: Write A PFS: Write B Professional Write 1 Professional Write 2 IBM Writing Assistant First Choice word procesor First Choice 3 word processor * PGL HP Graphics Language * PIC Lotus PIC * PICT Macintosh PICT Macintosh PICT2 * PNTG MacPaint * PP2 PowerPoint 3.0 for Windows PowerPoint 4.0 for Windows PowerPoint 4.0 for Macintosh * PP7 PowerPoint 7.0 for Windows 95 * PPL PFS: Plan * QA Q&A Write * QAD Q&A Database * QP6 Quattro Pro 5.0 for Windows Quattro Pro 6.0 for Windows Quattro Pro 7.0 for Windows * RBS R:Base System V R:Base 5000 * RFT IBM DCA/RFT * RFX Reflex * RTF Rich Text Format * SAM Samna * SC5 SuperCalc 5 * SYBASE * SDW Ami Draw * SHW3 Novell Presentations 3.0 * SGML * SMD Smart DataBase * SMS Smart Spreadsheet * SMT SmartWare II * SNAP Lotus Snapshot * SPT Sprint * TAZ UNIX Compress UNIX Tar * TEXT Text - DOS character set Text - ANSI character set Text - Macintosh character set Text - Unicode character set UUencode * TGA Targa * TIF6 Tagged Image File Format EPS (TIFF header only) CCITT Group 3 Fax CCITT Group 4 Fax JPEG JFIF (JPEG not in TIFF format) * Teradata * TW Total Word * TXT IBM DisplayWrite 2 or 3 IBM DisplayWrite 4 IBM DisplayWrite 5 * VW3 Volkswriter * W6 Microsoft Word 6.0 for Windows Microsoft Word 7.0 for Windows 95 Microsoft WordPad * WG2 Lotus 123 for OS/2 release 2 * WK4 Lotus 1-2-3 3.0 Lotus 1-2-3 4.0 Lotus 1-2-3 5.0 * WKS Lotus 1-2-3 1.0 Lotus 1-2-3 2.0 Lotus Domino/Notes Symphony Microsoft Works SS Microsoft Works DB VP-Planner Mosaic Twin Quattro (DOS) Quattro Pro (DOS) Generic WKS Windows Works Spreadsheet Windows Works Database * WM WordMarc * WMF Windows Metafile * WORD Word for Windows 1.x Word for Windows 2.0 Word for Macintosh 4.0 Word for Macintosh 5.0 * WORK Microsoft Works DOS 1.0 WP Microsoft Works DOS 2.0 WP Microsoft Works Win 3.0 WP Microsoft Works Win 4.0 WP * WP5 WordPerfect 5.x * WP6 WordPerfect 6.0 WordPerfect 6.1 WordPerfect 7.0 * WPF WordPerfect 4.2 * WPG WordPerfect Graphic 1.0 * WPG2 WordPerfect Graphic 2.0 WordPerfect Presentations * WPS Novell PerfectWorks 3.0 word processor CONVERA CONFIDENTIAL Novell PerfectWorks 3.0 draw Novell PerfectWorks 3.0

Chci mít nejlepší na začátku seznamu !!!

The image shows three overlapping web browser windows illustrating search results for the query "relevance ranking".

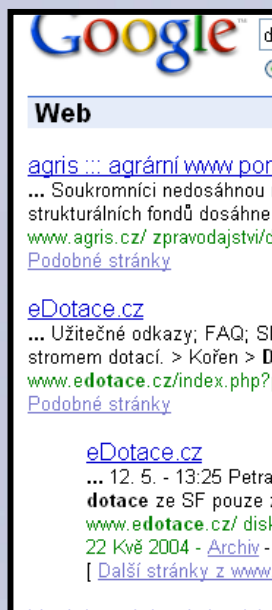
- Top Window (Verity, Inc.):** Shows search results for "relevance ranking" on the Verity website. The address bar contains: `Its1.jsp?query=relevance%20ranking&summary=no&Verity=Verity/¤tPage=1&Rec=10&sortField=Score&sortOrder=desc`. The page title is "Search Results".
- Middle Window (Google):** Shows search results for "relevance ranking" on Google. The address bar contains: `http://www.google.com/search?q=relevance ranking`. The search results include:
 - Web Results:**
 - [LT-World Result](http://www.lt-world.org/elf/elf/collatequery?h_...): Relevance Ranking Definition: Queries systems are often not very specific, and le...
 - [Relevance Ranking](http://www.contecds.com/library/c2/reirank.htm): Relevance Ranking. The most powerful is relevance ranking. Simply put, relev...
 - [Relevance Ranking](http://www.contecds.com/library/c2/reirank.htm): Relevance Ranking. Relevance Ranki by which C2 assigns significant values to...
 - [Dummies: Understanding Search F...](http://www.dummies.com/WileyCDA/Dummies...): Home > The Internet > Using the Internet
 - [Relevance Ranking. Understanding Sea...](http://www.dummies.com/WileyCDA/Dummies...)
 - [ACMT: Changing the relevance ra...](http://www.parkinsn.coles.org.uk/info/relevanceRanking.html): Changing the **relevance ranking** of certa the Termweight operator to change the re...
 - [Relevance Ranking](http://www.dummies.com/WileyCDA/Dummies...): Technologies. **Relevance Ranking**. Highly accurate natural language relevance
- Bottom Window (RetrievalWare Search):** Shows search results for "relevance ranking" on the RetrievalWare website. The address bar contains: `http://search.convera.com:8080/rwareJsps/RWRResults.jsp?SDOC=1`. The page title is "RetrievalWare 8". The search results include:
 - Search Results:** Search for: **relevance ranking** -- 25 document(s) match your query
 - Results List:**
 - (100) 355.pdf: Lima based newspaper Diario El Comercio is one of the oldest publications in Lima based newspaper Diario El Comercio is one of the oldest publications in the Americas. Established in the 1840's, today it is the largest circulation quality daily in Peru. El Comercio publishes 365 days a year, and is seen as the most authoritative
 - (100) <http://www.convera.com/press/docs/concept09.pdf>: 1. The key to a quality portal 3. Driven by content 5. Presenting a clear view 8. A universal solution 10. The Knowledge Refinery 13. Rabobank's self-service solution 15. More than the tip of the iceberg 18. Building up powerful collaboration 21. Sharing
 - (100) <http://www.convera.com/press/docs/concept10.pdf>: 1. From search and retrieval to knowledge discovery 3. Innovation as a competitive advantage 5. Unisys ? Virtual communities are now reality 8. The deterministic approach 10. Easy steps to great web site search 14. Speak local, think global 16. Exploiting
 - (75) 323.pdf: Maximizing Organizational ? Know How? in Government Entities
Maximizing Organizational ? Know How? in Government Entities Convera's New Generation of Knowledge Management



Potřebuji PŘESNOST A HODNOCENÍ VÝSLEDKŮ

Schopnost vyhledat **všechny**
relevantní dokumenty

Schopnost vyhledat **pouze** relevantní
dokumenty



2.	<input checked="" type="checkbox"/>	100%	Název: CR- Autor: CTH Datum: 13. Knihovna: Arc
3.	<input checked="" type="checkbox"/>	98%	Název: SO- Autor: CTH Datum: 13. Knihovna: Arc
4.	<input type="checkbox"/>	92%	Název: PL- Autor: CTH Datum: 104 Knihovna: Arc
5.	<input type="checkbox"/>	92%	Název: ČR- Autor: CTH Datum: 16. Knihovna: Arc



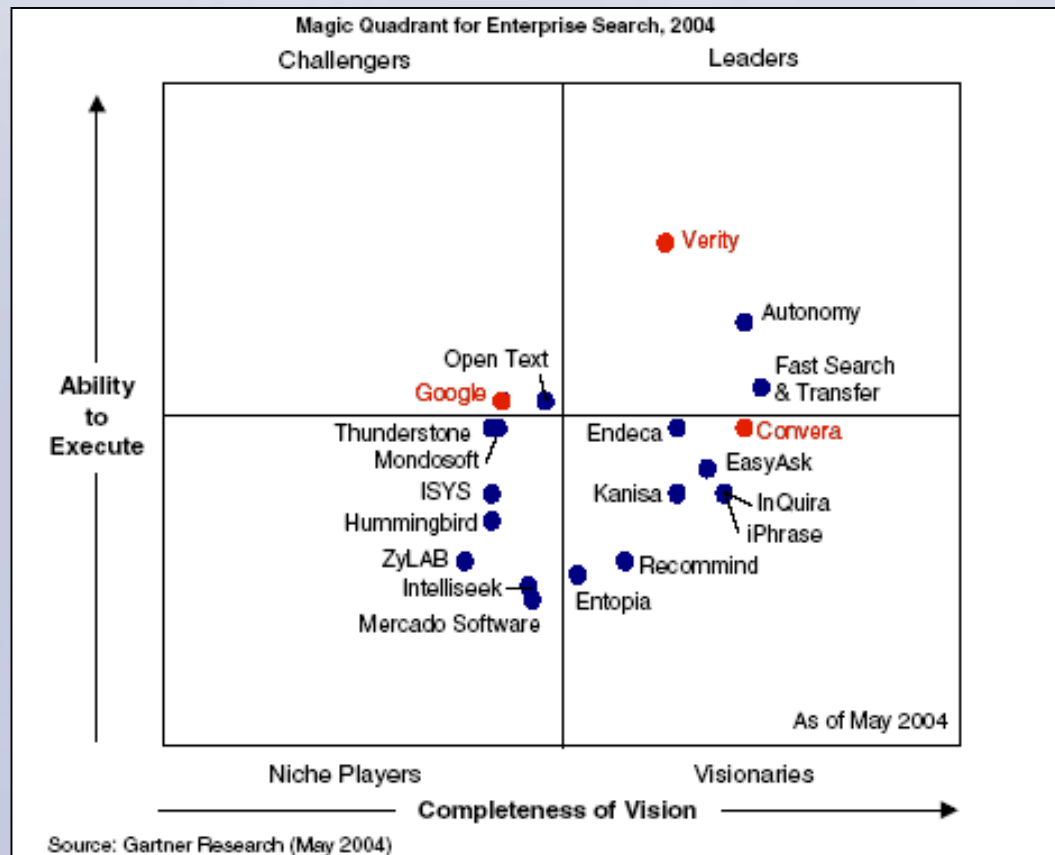
Kde se RR používá ?

- Vyhledávání
- Řazení vyhledaných dokumentů
- Kategorizace a klasifikace dokumentů

- Všechny formáty a typy informací



Na čem budeme vysvětlovat ...



Nejrozšířenější & Nejsilnější & Nejvyspělejší



příklad 1: **Google** (nejen internetový)

Google PageRank

Základní hodnocení relevance využívané systémem založené na analýze odkazů (hyperlinků) - obsažených v dokumentech a odkazujících na dokument a hodnotě odkazujících dokumentů

Dokumenty neobsahující odkazy nedosáhnou vysoké relevance



Příklad 2: Verity

VERITY RR

- **Základní vyhledávání**

největší relevanci mají **dokumenty obsahující nejvíce RŮZNÝCH klíčových slov** tvořících dotaz a jemněji pak ty dokumenty, ve kterých se tato **slova vyskytují s největší hustotou**

- **pojmovém vyhledávání**

vychází z váhy klíčových slov tvořících listy stromu. Pokud se slovo nebo jeho ekvivalent v textu vyskytne, přenáší se do nadřazeného uzlu přírůstek relevance dokumentu v závislosti na příslušném operátoru a hustotě výskytu slova (volitelně). Tak se postupuje až ke kořeni stromu.

- **expertním vyhledávání**

využívá adaptivní úpravy relevance v závislosti **na hodnocení dokumentu zvolenou skupinou uživatelů či authority** nebo v závislosti na osobním profilu.



příklad 3: **Convera**

CONVERA RW RR

Hodnocení ve dvou úrovních – pro vyhledání a pro uspořádání výsledků

- Úplnost
- Kontextová evidence
- Sémantická vzdálenost
- Blízkost
- Hustota nálezů
- ...

RR není závislé na statistikách výskytu jednotlivých slov v úložištích – výsledky lze hodnotit v rámci jednoho seznamu



Základní způsoby ...

Hodnocení relevance probíhá zpravidla ve dvou fázích

1. „vyhledání“ 3
2. „uspořádání výsledků do seznamu“ 2



Hodnocení

1. vyhledání

■ Úplnost

Čím je vyšší počet slov nalezených v dokumentu tím více je hodnocen

Dotaz: „Dotace z evropské unie“

Nalezeno: **Dotace z evropské unie**



2. Vyhledání

■ **Kontextová evidence**

Čím je vyšší počet termínů souvisejících významově v blízkosti nalezených slov z dotazu tím vyšší je hodnocení

Dotaz: „Dotace z evropské unie“

Nalezeno: **Financování** prostřednictvím **dotací**
.... Prostředky uvolňované ze strukturálních fondů **EU**“



3. Vyhledání

■ **Sémantická vzdálenost**

Čím je vyšší počet úzce souvisejících termínů ke slovům v dotazu tím vyšší je hodnocení

Dotaz: „Dotace z evropské unie“

Nalezeno: **Dotace > finance > prostředky**



4. Uspořádání

■ **Blízkost**

Dokumenty obsahující slova z dotazu a související slova pohromadě (věta, odstavec) mají vyšší hodnocení



5. Uspořádání

■ **Hustota nálezů**

Čím vyšší je poměr slov z dotazu a souvisejících slov k celkovému počtu slov v dokumentu tím je vyšší hodnocení



#. Vyhledávání a uspořádání

■ **Algoritmické**

Co s médii ?

Hodnoceni relevance založené pouze na statistikách podobnosti pro **VIDEO / ZVUK / IMG**

Např.:

Convera CST (colour, shape, texture)

Nexidia (sound shape)

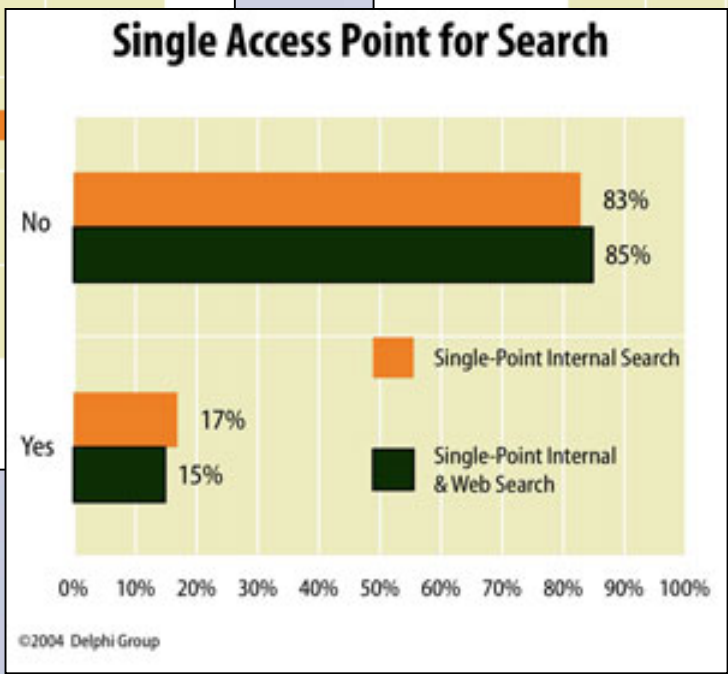
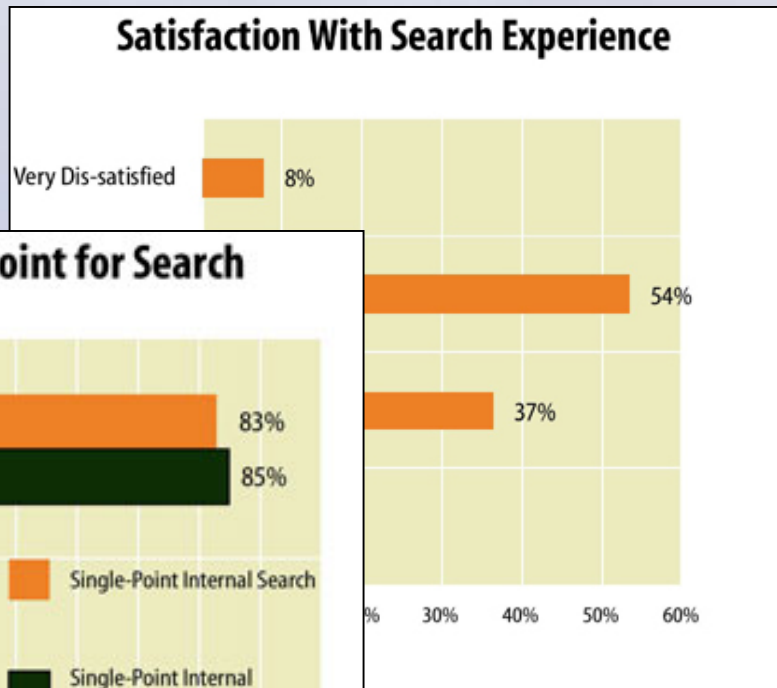
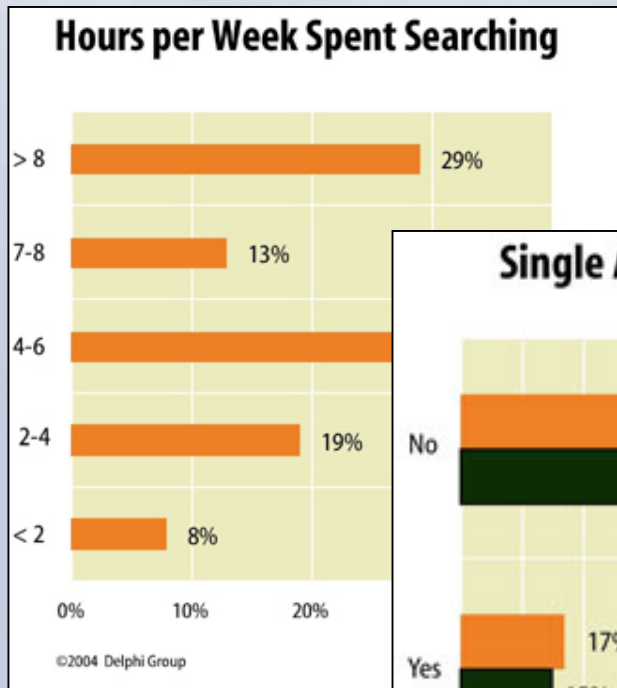
The screenshot shows the 'CST Image Retrieval' application window. It features a search bar with the keywords 'flowers plants' and a 'Max Hits' field set to 8. A table on the right lists various image features and their contribution to the score. Below the table, a grid of image thumbnails is displayed, each with its corresponding file name and similarity percentage.

	Contribution to Score	
Color Content		5
Texture Content		5
Grayscale/Shape		5
Color/Shape		5
Aspect Ratio		5
Keywords		5

Image	File Name	Similarity
	330017.BMP	100.0%
	330036.BMP	96.3%
	330034.BMP	93.6%
	330088.BMP	93.4%
	330045.BMP	92.7%
	330035.BMP	92.7%
	330055.BMP	92.7%
	330070.BMP	92.6%



Vyhledávání – DG 2004



Přesnost a efektivnost vyhledávacích řešení je stále důležitější ...



***“Without accuracy,
search engines are
working non-solutions”***

GartnerGroup, 2000



Dotazy ?

Děkuji za pozornost v toto pozdní odpoledne ☺

