VYSOKÁ CENA NEDOSTUPNOSTI INFORMACÍ

Dominik Mathauser
Pavel kocourek
KMA INCAD

UISK, FF UK

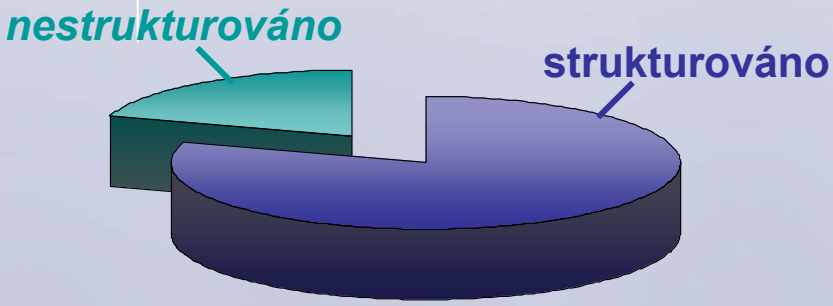
Současná situace

- Stále narůstající objem dat
- Rozprostření datových úložišť
- Velký počet formátů dokumentů
- Vysoké procento nestrukturovaných / neindexovaných dat

= „NEVIDITELNÉ“ INFORMACE

Hodnota informací

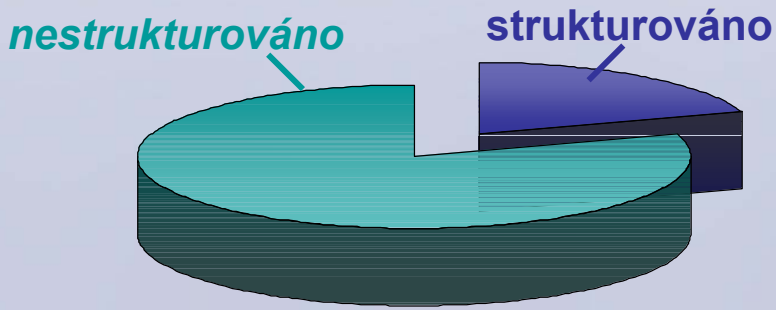
IT INVESTICE



Většina IT investic směřuje
k budování systémů
se strukturovanými informacemi

- **ERP**
- **DMS**
- **BI**
- **CRM**
- **ECM**
- **finanční IS**

OBCHODNÍ HODNOTA



Většina podnikových znalostí je
uložena v systémech pro
nestrukturované informace

- **Web servery**
- **Knowledge Management**
- **Content Management**
- **Groupware**
- **File Systemy …**

Source: Document and Content Technologies Market Forecast and Analysis, 2000–2004 IDC

Co způsobují „neviditelné“ informace

- Špatné rozhodování založené na nekompletních nebo dokonce chybných informacích
- Zbytečná práce - znovu-zpracování již existujících podkladů
- Snížená produktivita - zaměstnanec nemůže najít potřebné informace
- Ztráta obchodních příležitostí – zákazník nenajde požadovanou informaci

Scénář 1: čas ztracený hledáním

- Narůstá objem informací ve společnostech
- Roste počet úložišť

- Ne všechny informace ve společnostech jsou indexované nebo prohledávatelné

Scénář 1: čas ztracený hledáním

- Poměr neindexovaných informacích ve společnostech
  - 35% - 50%

  - **Předpoklad:**
    - Plat informačního pracovníka = 350 000 Kč / rok
    - Počet pracovníků = 500
    - Doba strávená hledáním denně = 2,5 hod (IDC)
    - Poměr neindexovaných = 50%

    - Závěr : společnost zaměstnávající 500 inf. Pracovníků ztrácí týdně **494 910 Kč** nebo **25 735 294 Kč** ročně.

Scénář 2: čas ztracený znovu-vytvářením informací

- Studie společnosti IDC (1999) zjistila, že 500 nejbohatších společností (Fortune 500) ztrácí ročně **12 miliard $** v důsledku:
  - opakovaného vytváření existujících řešení
  - podprůměrných výkonů
  - neschopností najít informační/znalostní zdroje

= ZNALOSTNÍ DEFICIT
(KNOWLEDGE DEFICIT)

Scénář 2: čas ztracený znovu-vytvářením informací

- Průměrný náklad na znalostní deficit v EU je 5 850$ ročně
- Průměrný náklad na znalostní deficit v ČR je 23 692 Kč (poměr ročního platu zaměstnance v ČR vůči „starým“ členům EU je cca 15%)

  - **Předpoklad :**
    - Počet informačních pracovníků = 500
    - Znalostní deficit v ČR = 23 692 Kč

    - Závěr: společnost zaměstnávající 500 inf. pracovníků ztrácí **11 846 000 Kč** ročně v důsledků vytváření již existujících informací

Scénář 3: promeškané příležitosti podniku

- promeškaný čas, který mohl být využit efektivněji
- nedostatek informací pro rozhodování

- **Předpoklad:**
  - Roční výnos podniku = 1 000 000 000 Kč
  - Počet informačních pracovníků = 500
  - Výnos na inf. pracovníka/rok = 2 000 000 Kč
  - Výnos na inf. pracovníka/hod = 900 Kč

  - Závěr: společnost ztrácí **620 040 Kč** týdně nebo **32 242 063 Kč** ročně v důsledku promeškání příležitostí.

- „Knowledge workers spend more time unwittingly recreating existing knowledge than in creating new knowledge“

Kit Sims Taylor – *International Conference on the Social Impact of Information Technologies* in St. Louis, Missouri, 1998

Co s tím

aneb jak vytvořit z informační záplavy informační bohatství…

- **platforma pro pokročilé znalostní vyhledávání a kategorizace**

Obsahové vyhledávání

- Rozeznává téma podle obsahu
- Expanduje uživatelské dotazy s použitím
  - gramatiky daného jazyka
  - znalostních bází (slovníků, tezaurů ...)
- Sémantické sítě

Přirozený jazyk

- Uživatel by neměl být nucen formulovat složité dotazy nebo používat operátory a zástupné znaky
- Vyhledávací nástroj musí být připraven ke zpracování přirozeného jazyka
- Jednoduchost používání je kritickým prvkem

Hodnocení relevance

- Nejlépe hodnocené dokumenty na počátku přehledu vyhledaných
- RR:
  - kompletnost
  - kontextová souvislost
  - sémantická vzdálenost
  - blízkost nálezů
  - nálezy vs velikost dokumentu
  - ……

Podobnostní vyhledávání

- **Nezáleží na sofistikovanosti vyhledávání jsou-li chyby či překlepy v dokumentech nebo v dotazu**

- **Pattern Recognition**

Podpora jazyků

- Závisí na prostředí, ve kterém je systém nasazen

- Uživatelé zpravidla vyžadují:
  - rodný jazyk
  - v oboru používaný jazyk a terminologie

Typy dokumentů a dat

- Textové procesory
- Tabulkové procesory
- Prezentace
- Grafika
- Internetové formáty (html, XML, pdf)
- RDBMS data
- DMS a Groupware
- ...

Škálovatelnost

- Narůstá počet jednotlivých dokumentů
- Narůstá počet zdrojů
- Narůstá počet uživatelů, kteří čekají rychlé, přesné a spolehlivé odezvy při vyhledávání

- rozložitelnost výkonu na CPU nebo fyzické servery

Bezpečnost

základní předpoklad pro sdílení informací je jejich zabezpečení ...

Budoucnost ES

Oblasti inovace definované OVUM

- Rozšířená kategorizace a podpora taxonomií
- Profilace a spolupráce
- Práce se strukturovanými a nestrukturovanými daty
- Podpora multi-jazyčnosti
- Visualizace

Dotazy ?

Děkuji za pozornost

dominik@incad.cz

10. iFORUM 2004

Dominik Mathauser

KMA INCAD

ÚISK, FF UK