AUTOMATICKÁ OBSAHOVÁ ANALÝZA DOKUMENTŮ JAKO JEDEN Z NÁSTROJŮ PRÁCE S INFORMACEMI

Zdeněk Jonák, Výzkumný ústav pedagogický Praha

Motto:
*„Budoucnost nepatří těm, kdo budou informace produkovat nebo distribuovat, ale těm, kdo ovládnou nástroje na jejich zpracování ve smysluplných souvislostech.“ (*P. Saffo)

Můj referát je orientován spíše na školní knihovny, které mají, přes řadu znaků společných s veřejnými knihovnami, sobě vlastní specifika. Práce školního knihovníka je více zaměřena na práci s obsahem textu než jen s jeho vyhledáním a poskytnutím čtenáři. Více na práci se znalostmi než s informacemi. Uživatelé školní knihovny, student či učitel, přicházejí do školní knihovny s přesvědčením, že knihovník má vhled do jeho aktuálních problémů, a že mu s nimi, počínaje fází vyhledávání až do konečného zpracování dokumentu či vyřešení problému, bude nápomocen. Školní knihovník nemůže své problémy řešit tím, že pošle své uživatele do veřejné knihovny. Proto vycházím spíše z potřeb praxe školní knihovny, když tvrdím, že v současné době je nanejvýš aktuální začít v knihovně **využívat analytických nástrojů, které umožňují zpracovávat informace do podoby znalostí.**

Chtěl bych tím navázat na koncepci Státní informační politiky ve vzdělávání, která v oblasti školství neznamená jen seznámení s dostupným komerčním software, ale vnesení ICT do všech oblastí společenského poznání, a tedy i do sfér, které byly dosud převážně doménou intuice či tvořivé fantasie. Obecně se však tato skutečnost týká i knihoven všech typů. Je třeba si všimnout signálů, ozývajících se na vnitrostátních fórech a konferencích, dříve než budou přicházet jako pokyn z EU.

Na letošním Infosu (http://wwwikaros.ff.cuni.cz/ikaros/2000/c05/infos2.htm) bylo možné např. vyslechnout výhrady k funkci knihovny jako vzdělavatelsko-osvětové půjčovny dokumentů. Takový typ nemá dle některých referujících a komentátorů výhledově v době Internetu co nabídnout. Jak se vyhnout pocitu zániku klasické knihovny spočívajícím:

- ve „fetišizaci“ knihy, ať již tištěné či elektronické,
- v dosavadní formě decentralizovaných a tudíž duplicitních fondů ohrožované centralizovanou virtuální knihovnou.

Fetišizace, fyzikalizace knihy a informace, orientace knihoven na službu – to jsou hlavní bolesti knihoven. „Knihovníci ztrácejí jistotu“ referuje Z. Uhlíř

„omezuje se jejich možnost něco podat“. Perspektiva knihoven není v jejich institucionální, ale v procesuální formě. Řešením není ani spoléhání na ukládání fondu elektronických knih, protože jde stále o práci s uzavřeným dokumentem a cílem knihovny zůstává quasireprodukce dokumentu a nikoliv jeho obsahová analýza a adresování konkrétnímu uživateli.
Všimněme si, jak do oblasti knihovnictví zasahují služby, zatím paralelní, které však se mohou stát rázem jejich konkurenty, nezatíženými jejich institucionální podobou**.**

Služby typu Anopress, produkty založené na software Verity s jejich možnostnmi:
- vyhledávání
- filtrování
- uspokojováním požadavků na vyžádání institucí
- fungujících 24 hodin denně
- vícejazyčné fultextové služby

Jednou možností je příklon knihoven k posílení badatelského přístupu. K tomuto cíli existují již dnes nástroje. Je třeba připustit, že nedokonalé, ale v době, kdy jejich potřeba v knihovnictví nazrává, je účelné jejich vývoj sledovat a respektovat. Jde o potřebu orientovat se v nepřehledném množství dokumentů, vyhledat dokumenty podobné obsahem a hledat možnosti, jak je od sebe automaticky kvalitativně odlišit a poskytnout uživateli jen ty obsahově blízké jeho potřebám.

Proč se obsahová analýza dostává v současné době v odborných kruzích v zahraničí do popředí zájmu? Dle mého názoru to souvisí i s tím, že se poptávka po technologických inovacích začíná nasycovat a do popředí vystupují potřeby, které byly realizací počítačového zázemí pozdrženy.

Mezi nasycením informačních institucí výpočetní technikou a zesílením zájmu o obsahové problémy klasických i elektronických médií existuje přímá vzájemná souvislost. Zvýšený zájem o analýzu obsahu pravděpodobně signalizuje další etapu rozvoje informačních služeb, spočívající ve využití výpočetní techniky, a to nikoliv jen k popisu a dopravě informačních zdrojů, ale i k jejich hlubšímu poznání, a tím i k splnění kritéria adresnosti informačních služeb uživateli. V souvislosti s Internetem je obtížné si například představit spojení mezi texty pomocí hyperlinků bez předchozí obsahové analýzy, která odhalí mezi miliony textů obsahové podobnosti a spojí je automaticky do podoby hypertextu. Na to dosavadní mocné, ale zatím lingvisticky jednoduché roboty nestačí.

Tento referát využívá výsledky teoretických úvah, které jsem na toto téma průběžně v předchozích letech publikoval v elektronickém časopise IKAROS (http://ikaros.cz)

Cesta od pořádání informací k pořádání znalostí

V době, kdy Internet poskytuje nepřeberné množství informací, se laickému uživateli informačních systémů mohou zdát problémy týkající se pořádání informací do vyšších struktur – poznatků, znalostí a teorií - málo smysluplné. Zdá se mu, že stačí jen usilovně hledat a hledat. Odborníci, vědečtí pracovníci, špičkoví manažeři, ale kupodivu i žáci a studenti, kteří potřebují často proměnit informace ve vědomosti, si však stále naléhavěji uvědomují, že právě přemíra a vzrůstající neuspořádanost informací vytváří řadu problémů, které je potřeba rychle řešit.

Při řešení těchto problémů se jeví jedním z možných východisek přístup, který zavádí do souborů informací určitou organizaci. Existuje instituce, která se těmito problémy řadu let zabývá a jejímž cílem je vyvést problematiku ukládání, zpracování a distribuce informací z úzkých hranic osobních, oborových či národních zájmů. Tato organizace se jmenuje International Society Knowledge Organisation (ISKO). Již několik let se snaží tato organizace navázat s naší zemí spolupráci. Přináším stručnou informaci o jejím poslání a aktivitách a snaze užšího okruhu lidi, kteří se snaží přenést aktivity této organizace i na naše území.

Co je ISKO a co rozumíme pořádáním znalostí?

ISKO - byla založena v r. 1989. Základním smyslem organizace je podporovat koncepční, teoretickou a metodickou práci v oblasti reprezentace znalostí ve všech formách.

V současnosti je členy ISKO více než 500 odborníků z oblasti informační vědy, filozofie, lingvistiky, informatiky a dalších disciplín, kromě jiného i ze specializovaných informačních oborů.

Pořádání znalostí (Knowledge organization) je pojem, který se v domácí literatuře zatím příliš nepoužívá, a proto neexistuje ani jeho obecně přijímaný český ekvivalent. Lze použít doslovného překladu (organizace znalostí), vhodnější však je použít termínu pořádání znalostí, který lépe vyjadřuje smysl pojmu. Volnější, nicméně velmi přijatelný je termín reprezentace znalostí.

Reprezentace znalostí je oblast lidského poznání, která se zabývá organizací jednotek znalostí (pojmy, teorie, hypotézy, poznatky, informace, data) a objektù všech typů, které odpovídají pojmům (teoriím...) nebo pojmovým třídám tak, aby byly zachyceny znalosti o světě a umožněno rozšiřování těchto znalostí pro účely jejich využití.

Mezi základní cíle organizace patří:

- mezioborová spolupráce odborníků z oblasti zpracování informací a znalostí, podpora výzkumu, rozvoje a aplikace systémů reprezentace znalostí, které rozvíjejí filozofické, psychologické a lingvistické přístupy k pořádání znalostí, poskytování komunikačních a síťových prostředkù knowledge organization pro členy ISKO

- spolupráce s organizacemi, které se zabývají problémy spojenými se zpracováním informací a znalostí

V současné době se realizuje snaha vytvořit českou odbočku ISKO s možností implementovat metody a přístupy této mezinárodní organizace v českých podmínkách. Zájemci mohou získat podrobnější informace na www stránce : http://ikaros.ff.cuni.cz/ikaros/1999/c01/isko.htm

Do oblasti ISKO patří i některé problémy, kterými se zabývám v tomto článku. Jedním z nejčastějších problémů je schopnost dosažení hlubšího porozumění textu. To předpokládá důkladnější přípravu textu před vstupem do databází či Internetu. Technologie zpracování se teprve rodí. Dnes se, bohužel, v úvahách o obsahovém zpracování dokumentů směšují dva přístupy:

- systémy zpracování numerických dat
- systémy zpracování textových dat

Oba typy informačních systémů se často zastřešují společným názvem „data mining“. Ztotožnění obou způsobů zpracování dat je velmi zkreslující, protože jde o přístupy z hlediska nároků na vytvoření algoritmu a programového řešení, co do obtížnosti, nesrovnatelné. V našem referátu se zaměřujeme pouze na systémy analýzy textových souborů.

Co je pro textové systémy charakteristické?

Dokumenty z obecného hlediska představují prostředky přenosu či uchování modelů skutečnosti ve znakové podobě. Mezi modely a soubory znaků, které je reprezentují, existuje určitý stupeň volnosti, projevující se tím, že rozdílné modely skutečnosti lze popsat soubory znaků s vysokým počtem shodných prvků i vazeb mezi prvky a naopak k popisu obsahově blízkých modelů lze použít soubor znaků s velmi vysokým počtem rozdílných prvků a jejich vazeb.
Rovněž dotaz položený vyhledávacímu systému je dotazem tazatele po modelu určité skutečnosti. Od plného textu se liší především počtem slov. Tato redukce počtu slov není pro systém, který má dotaz zpracovat, žádnou výhodou. Subsystém zpracování dotazu, má-li být vyhledávání skutečně účinné - tzn., má-li získat ze souboru textů nabízejících formální podobné dokumenty, dokumenty shodné obsahově, musí vykonat řadu intelektuálně náročných operací.

Informace versus znalosti

V souvislosti s užíváním pojmů fakta, data, informace, znalosti, poznatky existuje mnoho nejasností. Obvykle se mezi pojmy informace, znalosti nediferencuje. V odborné literatuře se někdy setkáme s názorem, že znalosti jsou obecnější pojem než informace, jindy je tomu naopak. Dosavadní přístupy k analýze textů – internetové roboty, které analyzují text na jednotlivá slova ani hlubší vhled do struktury textu neumožňují.

Současné nástroje automatické analýzy, které chci představit, umožňují pracovat s nejbližšími vyššími celky textu, a to s KATEGORIEMI. Kategorie

musí na základě lexikálního rozboru textu vytvořeného analytickým systémem ovšem zatím vytvořit člověk, aby je systém obsahové analýzy potom sám smysluplně využíval.

Pro ilustraci nabídnu v následujícím textu dva systémy, první, který může sloužit k jakési přípravě neuspořádaných textových, obrazových, zvukových souborů do formy použitelné pro analýzu, druhý, který umožňuje ze souboru textů i textově popsaných obrazových a zvukových souborů vytvořit obsahové kategorie, jako materiál pro obsahovou analýzu souborů textů.

V době růstu množství neuspořádaných informací se považují za aktuální dva kroky:

- **Úprava neuspořádaných textových, obrazových, zvukových souborů do formy použitelné pro analýzu,**
- **Extrahování obsahových kategorií z textu, jako podkladový materiál pro obsahovou analýzu souborů textů.**

ad 1. Úprava neuspořádaných textových, obrazových, zvukových souborů do formy použitelné pro analýzu, ATLAS.ti

Cílem systému ATLAS.ti je racionalizovat práci s datovými soubory obsahujícími texty, obrázky či zvuky, vizualizovat jejich strukturu a zefektivnit tak jejich vyhledávání jako předpoklad pro obsahová zobecnění či tvorbu teorií. Systém není přirozeně dokonalý. Neřešena zůstává problematika synonymie, homonymie, akceptování významu větných a nadvětných celků. K řešení těchto nedostatků však poskytuje systém vynalézavé pomůcky. V každém případě je však snaha vnést řád do struktury datových souborů jednotlivého uživatele či instituce, předtím, než se dostanou do sítě WWW jedním z předpokladů k dosažení vyšší úrovně jejich dalšího využívání.

- Vytvářet ze všech dokumentů, klíčových slov a odkazů tzv. hermeneutickou jednotku, jejímž cílem je vzájemně provázat identifikovatelné prvky datových souborů,
- usnadnit vyhledávání, segmentaci, spojování prvků do rodin dle podobnosti, porovnání textů. Vytvářet předpoklady pro aplikace dalších statistických a lingvistických metod,
- podporovat tvorbu WWW stránek.

Pojem „hermeneutická jednotka“ již sám mnohé napovídá o cíli systému. Hermeneutika je věda, usilující o pravdivé, věrohodné, hlubší pochopení textů. Zabývá se tedy procesem interpretace. Systém Atlas.ti si neklade za cíl činnost tak náročnou jako je interpretace textu, ale snaží se poskytnout nástroje, které strukturují nejasné vícevýznamové prvky textu jasnější charakteristikou, prováží obsahově příbuzné části textu vazbami a spojí tyto prvky do vyšší jasně identifikovatelné kategorie - hermeneutické jednotky. Systém pracuje ve dvou modech: textovém a pojmovém.

Textový režim realizuje segmentování datových souborů na dílčí obsahové úseky, označování textů, obrázků a zvuků. Uvedeným prvkům lze přiřadit vlastní indexy, anotace, komentáře. Pojmový režim spojuje vytvořené segmenty a znaky do sémantických sítí a umožňuje tak jejich vizualizaci a přehlednost. Tyto činnosti jsou předpokladem pro rychlé a bezeztrátové vyhledávání v datových bázích.

Schéma postupu od prvků textu k hermeneutické jednotce:
Surovinou je primární soubor, kterému se přiřadí seriálové číslo jako jeho identifikační znak. Prvky systému představují následující jednotky:
Quotations - segmenty textu obsahující relevantní informace z textu
Codes - označení přiřazená segmentům
Families - propojení prvků označených kódy (textů, obrázků, zvuků) do sémantické sítě

K vytvoření hermeneutické jednotky slouží: Editor pro tvorbu hermeneutické jednotky sestává z levého a pravého okna.V levém okně je uložen text, obrázek, zvukový soubor. V pravém okně se umisťují výše zmíněné prvky hermeneutické jednotky, představující jakási klíčová slova, umožňující charakterizovat segment textu, obrázek či jeho část a zvuk, k němuž se vztahují.
Network editor, umožňující propojit klíčová slova a segmenty vzájemnými vazbami. Relation editor - poskytuje vazebné operátory pro network editor. Např. is part of, is cause of, is property of.

Kódování obrázků a zvuků. Práce s obrázky a zvuky je podobná práci s texty. Tažením označíme část obrázku či zvukový úsek a označenému výseku přiřadíme klíčové slovo. Díky této vlastnosti je systém prakticky využitelný v řadě oborů, kde dochází k provázání těchto oblastí: v medicíně, výtvarném umění, architektuře, grafologii, kriminologii apod.

Výsledkem intelektuálně náročné práce se systémem ATLAS.ti je sématicky provázaný systém dokumentů, obrázků a zvukových souborů poskytující vizuálně přehlednou síť vztahů ve struktuře textů jako podklad další analýzy.

Ad. 2 **Extrahování obsahových kategorií z textu, jako podkladový materiál pro obsahovou analýzu souborů textů**

Obsahová analýza dokumentu představuje vedle řady rutinních knihovnických procesů činnost, vyžadující nejen mnoho času, ale i vysoký podíl intelektuální a tvůrčí činnosti. V důsledku toho se pojem obsahové analýzy v knihovnách omezil na proces letmé prohlídky dokumentu a jeho několikařádkový popis. Takto zúžený pohled na obsahovou analýzu vůbec neodpovídá představě autorů, kteří metodu obsahové analýzy uvedli někdy v první třetině 20. století v život. (Záměrně neodkazuji na tisíciletou tradici obsahové analýzy, za kterou lze označit studium bible apod.)

Obsahová analýza v pravém slova smyslu je metodou, která zmapováním kvantitativní struktury souboru textů dospívá k tomu, co jednotliví autoři,

v důsledku své subjektivní zaujatosti, neměli a ani nemohli mít v úmyslu sdělit - k odhalení latentního obsahu dokumentů. V souboru textů je vždy uloženo větší množství informací, které nelze prakticky získat čtením jednotlivých textů, poněvadž čtenář či analytik jsou schopni srovnávat jednotlivé texty zpravidla jen z jednoho hlediska. Obsahová analýza zpracovávající obrovské soubory textů z velikého počtu aspektů může, díky využití počítačových kapacit a rychlostí, tuto úlohu plnit spolehlivěji. Po potřebě tohoto druhu informací sílí tlak. Prognostikové, podnikatelé, ekonomové se zajímají stále častěji o informace, které nejsou všeobecně dostupné a které by chtěli znát včas především pouze oni.

Představím systém, na kterém chci ilustrovat možnosti obsahové analýzy.
TEXTQUEST (http://www.intext.de/TEXTANAE.HTM)

Systém TextQuest je prozatím rovněž lingvisticky jednoduchý a nejnáročnější intelektuální práci za něho musí odvést člověk, ale mohu odkázat na další vývojové varianty, které jsou již na trhu, i když zpravidla zatím finančně nedostupné. Cílem článku je především ukázat, jaké prvky textu jsou k automatické obsahové analýze potřeba a jaké algoritmy je k jejich zpracování nutné vytvořit.

Z výsledků obsahové analýzy je možné se o knihovním fondu, za předpokladu, že je digitálně zpracován, dozvědět mnohonásobně více než poskytuje jakkoliv pečlivě zpracovaný katalog nebo bibliografie. Poukáži na adresu jednoho výsledku obsahové analýzy z rozhlasového a televizního vysílání, protože rozhlas či televize jsou vlastně knihovny audiálních či vizuálních záznamů.
Doufám, že již první pohled na pracnost způsobu zpracování čtenáře-knihovníky neodradí. (http://blisty.internet.cz/9906/19990624d.html).
TextQuest vytváří následující výstupy:
- slovník jednoslovných výrazů
- slovník slovních sekvencí
- slovník permutací slovních výrazů
- slovník konkordancí

Tyto slovníky mohou sloužit k jednoduchému popisu textů, jako jeho indexy nebo jako podklad k tvorbě kategorií pro obsahovou analýzu, analýzu obtížnosti textu a další zjištění.
Slovník jednoslovných výrazů je abecedně uspořádaný seznam výrazů s údajem o počtu výskytů.
Slovník slovních sekvencí je seznam sekvencí obsahující 2 - 4 slova.
Např. z věty: **Vybavení škol a knihoven informačními a komunikačními technologiemi**.

2 slova	**3 slova**	**4 slova**
Vybavení škol	*Vybavení škol informačními*	*Vybavení škol informačními a*
škol informačními	*škol informačními a*	*škol informačními*
		a komunikačními

*(*pozn.*:* předložky, spojky a další výrazy mohou být z textů eliminovány pomocí tzv. slovníku zakázaných výrazů). Potom by písmeno „a „ v uvedených sekvencích nebylo použito).

Slovník permutovaných slovních výrazů

Permutace představuje skupiny slov, vzniklé záměnou pořadí prvků slov dané množiny slov.

Vybavení škol
Vybavení a
Vybavení knihoven
Vybavení informačními
Vybavení a
Vybavení komunikačními

Slovník konkordancí

Konkordance slouží k zobrazení slova v kontextu spolu s údajem o umístění slova v kontextu. Analyzované slovo je umístěno uprostřed řádky a ostatní slova ho obtékají. Délku kontextu lze nadefinovat.

Ukázka:

je považován za	***básníka*** *rodného kraje*
kdybychom	***brali*** *v úvahu*
sbírce sleduje	***Březina*** *jakoby*

Využití výsledků obsahové analýzy

Porovnávání slovníků

Slovníky jednoslovných či víceslovných výrazů jednotlivých textů mají velkou vypovídací schopnost. Výsledkem je zjištění podobnosti/rozdílnosti slovníků. Na základě údaje o míře podobnosti/rozdílnosti, lze vytvářet hypotézy o obsahové podobnosti/rozdílnosti dokumentů a vytvářet podklady pro tvorbu algoritmů.

Lze například vytvořit program vytvořený na základě hypotézy, že texty jejichž slovníky obsahují větší počet shodných slov jsou obsahově podobnější. Výstup ve formě slovníků jednoslovných či víceslovných výrazů mohou sloužit jako prvky pro tvorbu selekčních jazyků, indexů apod.

Systém tvorby kategorií

Slovníky však lze využít k daleko náročnější a propracovanější analýze obsahu. Je ovšem potřeba vytvořit intelektuálně příslušné obsahové kategorie a uvést je do vztahu se slovníky textů.

Subsystém názvů kategorií

Struktura názvu (z oblasti sportu):

kód	*název (60 znaků)*
1	*části těla*
2	*druh sportu*
3	*sportovní náčiní*

Subsystém interaktivního přiřazování slovních výrazů jednotlivým kategoriím

Vyžaduje velké intelektuální úsilí při formulaci kategorií a při přiřazování slovních výrazů těmto kategoriím.

Slovník	Název kategorie
atom	*1. části těla*
auto	*2. druh sportu*
činka	*3. sportovní náčiní*

Využití výsledků systému TextQuest

Ve srovnání se schopnostmi lidského intelektu jsou výsledky analýzy systému TextQuest samozřejmě nedostatečné. Člověk dospěje po přečtení textu okamžitě k určitějším a přesvědčivějším poznatkům o obsahu než popisovaný systém. Na rozdíl od člověka může však systém číst po libovolnou dobu díla uložená v digitální formě, nikdy se neunaví a všechny je posoudí z libovolného počtu hledisek.

Výsledkem automatické obsahové analýzy je tabulka kategorií uspořádaných podle vah výskytů jednotlivých výrazů v jednom textu nebo souboru textů.

kategorie	výskyt	$T_1, T_2 \ \ldots \ T_n$	T_{celkem}
*Kategorie **A***	*hodnota*	*hodnota*	*hodnota*
*Kategorie **Z***	*hodnota*	*hodnota*	*hodnota*

Tabulku výskytů kategorií jednoho textu poskytující informaci o jeho obsahové struktuře lze např. využít pro popis textu ve vyhledávacích systémech

Tabulka výskytů souboru textů je podkladem pro obsahovou analýzu. Obsahovou analýzou zde rozumíme analýzu podobnosti obsahových struktur souborů textů. Výsledkem je možnost výzkumu vývoje sledované problematiky v časových řadách, vytvářet na základě zjištění údajů z minulosti odhady do budoucnosti apod.

Z knihovnického hlediska je důležitá možnost uspokojovat velice specifické a podrobné požadavky uživatele. Uživatel se již nemusí uspokojit s vyhledáním nepřehledného množství dokumentů na svůj dotaz, ale může získat, pokud formuluje přesně svůj dotaz, odpověď na to, jak se formulovaný problém vyvíjel, může vyjmout z rozsáhlejších monografií pouze tu část, která se zabývá jeho problémem důkladněji apod.

Kromě podkladů pro obsahovou analýzu poskytuje systém TextQuest podklady pro velmi zajímavá praktická využití: Měření čtivosti, čtenářské obtížnosti textu, využitelný zejména v pedagogické praxi při tvorbě učebnic a učebních textů odstupňovaných dle didaktické náročnosti pro jednotlivé stupně.

Čtivost textu

Problematika čtivosti je ve veřejném knihovnictví prakticky málo využitelná. Ve školství, zejména v nakladatelské činnosti učebnic má rozhodující význam při stanovení rozhodování, jaký stupeň obtížnosti jazyka zvolit pro určitý věkový stupeň žáka.

Měření jazykové obtížnosti textu. Čeština má z hlediska měření čtivosti textu ve srovnání s angličtinou specifické zvláštnosti. Proto nejsou anglické míry čtivosti vždy použitelné. V naší praxi je například znám tzv. Mistríkův vzorec, který měří:

- délku věty ve slovech
- délku slova ve slabikách
- počet všech slov
- počet všech různých slov

Při měření obtížnosti textu se vychází se z hypotézy:

- čím je věta delší, tím složitější je model skutečnosti, předkládaný čtenáři,
- čím je slovo delší, tím obtížněji je čtenář vnímá (délka též svědčí o frekvenci užívání slova. Většina frekventovanějších slov se díky komunikačním mechanismům zkracuje.
- počet rozdílných slov: rozdílná slova ztěžují čtenáři text tím, že častěji naráží na slova, která dosud v předchozím textu nevyskytla.

Systém TextQest používá následující algoritmus

Subsystém měření čtivosti umožňuje používat 8 různých vzorců čtivosti, založených na měření syntaktických kriterií textu. Oproti jiným vzorcům nepracuje se vzorcem 100 slov, ale s celým textem či jeho částmi.

Hodnota vzorce je mezi 0-100. Nejvyšší hodnoty dosahují texty nejlépe čtivé.

Jako ukázku uvádíme nejznámější – Flescheho vzorec:

$$REI = 206.835 - \left(\frac{\text{počet slabik}}{\text{počet slov}} * 0.864\right) - \left(\frac{\text{počet slov}}{\text{počet vět}}\right)$$

Závěr

Naplnit vizi společnosti znalostí, neznamená jenom dosažení toho, že každému bude stát na stole počítač, ale především dosažení schopnosti eliminovat z gigabytů informací ty, jimiž stojí za to se zabývat, investovat do systémů, které nahradí potřebu číst a analyzovat tytéž soubory dokumentů, tisící uživateli každým z jednoho jediného hlediska, systémem analýzy obsahu textu jednou provždy.

Tyto požadavky nelze realizovat sebelépe organizovanou manuální ani intelektuální cestou, ale pouze za přispění technologií, jejichž tvorby se účastní lingvisté, programátoři, kognitivní vědci a patrně i psychologové a sociologové.