Slovenská technická univerzita v Bratislave

FAKULTA INFORMATIKY A INFORMAČNÝCH TECHNOLÓGIÍ

FIIT-5212-8390

Ľubomír Vnenk

Prehliadanie informačného priestoru s využitím kontextu aktivity

Bakalárska práca

Študijný program: Informatika
Študijný odbor: 9.2.1 Informatika
Miesto vypracovania: Ústav informatiky a softvérového inžinierstva, FIIT STU Bratislava
Vedúci práce: prof. Ing. Mária Bieliková, PhD.

máj, 2014

Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií

ZADANIE BAKALÁRSKEHO PROJEKTU

Meno študenta:	**Vnenk Ľubomír**
Študijný odbor:	Informatika
Študijný program:	Informatika
Názov projektu:	**Prehliadanie informačného priestoru s využitím kontextu aktivity**

Zadanie:
Informačné potreby sa v súčasnosti najčastejšie definujú kľúčovými slovami, na základe ktorých sa identifikujú dokumenty, ktoré môžu byť potenciálne zaujímavé pre používateľa. Efektívna práca s kľúčovými slovami pri navigácii v informačnom priestore vyžaduje skúseného používateľa, ktorý vie dobre identifikovať dôležité slová z dokumentov a mapovať ich na svoje ciele. Toto je veľmi obtiažne najmä v prípadoch, keď nemáme presne stanovený cieľ prehliadania informačného priestoru. Pomôcť môže kontext prehliadania informačného priestoru.
Skúmajte možnosti využitia kontextu pri navigácii v informačnom priestore. Zamerajte sa pri tom na kontext aktivity určený aktuálnou činnosťou používateľa, ktorá viedla k inicializácii procesu napĺňania informačnej potreby. Navrhnite metódu, ktorou identifikujete kontext aktivity pre vybrané scenáre (napr. písanie dokumentu). Vytvorený kontext použite na obohatenie dopytu a vyhľadanie vhodných informačných zdrojov, napr. formou odporúčaných dokumentov. Navrhnuté riešenie implementujte vytvorením softvérového prototypu v prostredí digitálnej knižnice integráciou do systému Annota. Experimentujte s vlastnosťami riešenia a vyhodnoťte ho.

Práca musí obsahovať:

Anotáciu v slovenskom a anglickom jazyku
Analýzu problému
Opis riešenia
Zhodnotenie
Technickú dokumentáciu
Zoznam použitej literatúry
Elektronické médium obsahujúce vytvorený produkt spolu s dokumentáciou

Miesto vypracovania: Ústav informatiky a softvérového inžinierstva, FIIT STU, Bratislava
Vedúci projektu: prof. Ing. Mária Bieliková, PhD.

Termín odovzdania práce v zimnom semestri: 10. 12. 2013
Termín odovzdania práce v letnom semestri: 12. 5. 2014

Bratislava 23. 9. 2013

prof. Ing. Pavol Návrat, PhD.
riaditeľ ÚISI

Anotácia

Fakulta informatiky a informačných technológií

Slovenská technická univerzita v Bratislave

Autor: Ľubomír Vnenk
Študijný program: Informatika
Bakalárska práca: Prehliadanie informačného priestoru s využitím kontextu aktivity
Vedúci bakalárskej práce: prof. Ing. Mária Bieliková, PhD.
Máj, 2014

Vyhľadávanie špecifickej informácie je z dôvodu informačného preťaženia čoraz zložitejšie a namáhavejšie, pretože existuje veľké množstvo rôznorodých informácií. Zadávanie presného dopytu je preto kľúčové. Priemerný dopyt však obsahuje dva až tri kľúčové slová, ktoré nemusia vystihovať používateľov zámer dostatočne presne. Predpokladáme, že tieto kľúčové slová súvisia vo väčšine prípadov s používateľovou aktivitou. Navrhli sme preto metódu na obohacovanie dopytu pomocou kontextu aktivity. Kontext aktivity je reprezentovaný kľúčovými slovami získanými pozorovaním používateľa za posledný časový úsek v prostredí počítača. Navrhli a implementovali sme metódu, ktorá nájde prepojenie medzi kontextom aktivity v niektorej aplikácií a dopytom. Nie všetky aplikácie však súvisia s dopytom, preto sme navrhli metódu, ktorá určí na základe rôznych indikátorov aplikáciu, ktorej kontext aktivity je rovnaký ako kontext vyhľadávania. Dopyt následne obohacujeme na základe kontextu aktivity v tejto aplikácií. Metódu sme overili na malej skupine študentov, ktorých aktivitu sme zaznamenávali pomocou vytvorených modulov. Porovnávali sme ich explicitné označovanie aplikácie, ktorá súvisí s dopytom a našu metódu na určenie aplikácie, ktorá súvisí s dopytom.

Annotation

Faculty of Informatics and Information Technology

Slovak University of Technology

Name: Ľubomír Vnenk
Course: Informatics
Bachelor thesis: Context-aware Information Space Browsing
Supervisor: prof. Ing. Mária Bieliková, PhD.
2014, May

The Web has so much variable information, therefore searching a specific one is complicated. To find something valuable, specifying good query is crucial. However, average query consists of only about two words, which cannot specify intent of a searcher well. We suppose that these words follow in most cases the searcher's activity. We propose an approach for search query extension by activity context. The activity context contains words reflecting the searcher's activity context expressed by keywords gathered from one of the very recent activity provided by the searcher. We present a method for finding out searcher's activity context by logging content and his interaction between applications considering both standalone applications and applications running inside a browser. We have designed a method to use the logs to find a connection between the query and specific application. Then we extend the query by terms gathered by analyzing the selected application's content. We evaluate our approach in a series of experiments based on data gathered by monitoring small group of searchers by means of developed logger prototype. We compare explicit selected connections to connections found by our methods.

Poďakovanie

Chcem sa poďakovať vedúcej mojej práce, pani profesorke Márií Bielikovej za jej rady, skúsenosti a vynaložený čas, ktorými mi výrazne pomohla pri písaní tejto práce.

Ďalej sa chcem poďakovať Annota tímu za možnosť spolupracovať na projekte a celej PeWe skupine za konštruktívne diskusie, ktoré mi často ukázali nový smer a upozornili ma na možné problémy

V neposlednom rade sa chcem poďakovať mojej rodine a priateľom, ktorý ma podporovali počas celého môjho štúdia

Ľubomír Vnenk

Čestné prehlásenie

Čestne prehlasujem, že záverečnú prácu som vypracoval samostatne s použitím uvedenej literatúry a na základe svojich vedomostí a znalostí.

Ľubomír Vnenk

Obsah

1 ÚVOD 1

2 VYHĽADÁVANIE 3

2.1 Pokročilé vyhľadávanie 4

2.2 Obohatenie dopytu 4

2.3 Personalizované vyhľadávanie 6

2.3.1 Personalizácia na základe profilu používateľa 6

2.3.1.1. Personalizácia na základe histórie vyhľadávania 6

2.3.1.2. Personalizácia na základe používateľových záujmov 6

2.3.2 Kolaboratívna personalizácia 7

2.3.3 Porovnanie 7

2.4 Zistenie kvality vyhľadávača 7

3 KONTEXT VYHĽADÁVANIA 9

3.1 Kontext aktivity 9

3.2 Zohľadnenie kontextu 10

3.3 Získavanie informácií od používateľa 11

3.3.1 Explicitná spätná väzba 11

3.3.2 Implicitná spätná väzba 12

3.4 Existujúce kontext zohľadňujúce vyhľadávania 13

3.5 Diskusia 15

4 NÁVRH METÓDY OBOHACOVANIA DOPYTU 16

4.1 Získavanie kontextu aktivity používateľa 17

4.1.1 Zachytenie aktivity používateľa 17

4.1.2 Spracovanie záznamu aktivity 18

4.2 Nájdenie prepojenia medzi aplikáciou a dopytom 19

4.3 Obohatenie dopytu 23

4.3.1 Relevancia kľúčového slova vo vzťahu k aplikácii 23

4.3.2 Relevancia kľúčového slova vzhľadom k dopytu 25

4.4 Diskusia 26

5 REALIZÁCIA NAVRHNUTEJ METÓDY 28

5.1 Zaznamenávanie aktivity používateľa 28

5.1.1 Tabber 29

5.1.2 Annota-extension 31

5.1.3 Wordik 32

5.1.4 Moduly pre spracovanie prirodzeného jazyka 33

5.2 Analýza prepojení medzi dopytom a aplikáciou 33

6 EXPERIMENTÁLNE VYHODNOTENIE 35

6.1 Existuje prepojenie medzi dopytom a aplikáciou 36

6.2 Určenie aplikácie súvisiacej s dopytom 39

7 ZÁVER 44

LITERATÚRA 46

PRÍLOHY 49

A TECHNICKÁ DOKUMENTÁCIA

B POUŽÍVATEĽSKÁ PRÍRUČKA

C INŠTALAČNÁ PRÍRUČKA

D PRÍSPEVOK PUBLIKOVANÝ NA KONFERENCII IIT.SRC 2014

E PREDBEŽNÁ VERZIA PRÍSPEVKU PRIPRAVOVANÉHO NA KONFERENCIU ENIC 2014

F OBSAH ELEKTRONICKÉHO MÉDIA

1 ÚVOD

Internet sa stal najjednoduchším a najpoužívanejším prístupom k informáciám. Prehliadať však také obrovské množstvo informácií je zložité a vzhľadom k rôznorodosti informácií je zložité nájsť informáciu, ktorú potrebuje práve konkrétny používateľ. Toto je dôsledkom aj toho, že každý človek si pod daným kľúčovým slovom a dopytom môže predstavovať niečo iné. Vyhľadávač však nevie presne určiť, čo je cieľom používateľovho vyhľadávania, pretože má o používateľovi veľmi málo informácií. Nevie zistiť, čo ho viedlo k aktuálnemu vyhľadávaniu, čomu sa momentálne venuje a teda nemá veľa cenných informácií, z ktorých by usúdil cieľ aktuálneho vyhľadávania. Má len málo informácií na to, aby prispôsobil výsledky vyhľadávania konkrétnemu používateľovi, ktorý je jedinečný a má špecifické aktuálne záujmy a aktivity, pričom zadáva viacznačné dopyty.

Používateľ trávi na počítači veľa času, predovšetkým pred začatím vyhľadávania, preto je veľká pravdepodobnosť, že to, čo vyhľadáva, súvisí práve s niektorou aktivitou, ktorú pred chvíľou vykonával na svojom počítači. V našej práci sa snažíme zistiť, aké aktivity na počítači vykonával a či niektorá aktivita súvisí s vyhľadávaním. Ak totiž nájdeme aplikáciu, ktorá súvisí s dopytom, je vysoko pravdepodobné, že kontext tejto aplikácií je podobný kontextu dopytu. Ak je napríklad kontextom aplikácie písanie vedeckej štúdie o zápaloch oka, pri dopyte v ktorom sa vyskytuje slovo oko môžeme predpokladať, že dopyt súvisí s touto aplikáciou a teda používateľ ma v pláne vyhľadávať súvislosti s ľudským okom, nie morským, pretože to vyplýva z kontextu aplikácie. Je dôležité preto vedieť presne zaznamenať kontext aplikácií. Našou snahou je určovať kontext všetkých aplikácií, teda webových aj desktopových, za pomoci pozorovania používateľovho správania v aplikáciách.

Cieľom našej práce je navrhnúť, zrealizovať a overiť metódu získavania kontextu vyhľadávania na základe aktivity používateľa pri práci s aplikáciami. Snažíme sa nájsť súvislosť medzi dopytom a kontextom niektorej z aplikácií, na základe ktorej budeme vedieť určiť, ktorá aplikácia súvisí s dopytom. Súvislosť hľadáme porovnaním kľúčových slov dopytu a kľúčových slov kontextu aplikácie, pričom hľadáme medzi nimi syntaktickú podobnosť alebo sémantickú blízkosť. V prípade, že sa nájde súvislosť s viacerými

aplikáciami, tú správnu získame skúmaním rôznych indikátorov, ako napríklad aktívna dĺžka používateľa v danej aplikácií. Za aplikáciu, ktorá je prepojená s dopytom určíme tú, ktorá získa najvyššiu relevanciu zohľadnením týchto indikátorov. Získaný kontext použijeme pri rozšírení dopytu vyhľadávania.

V kapitole 2 analyzujeme rôzne typy vyhľadávania a metódy, ktorými sa dá jednoduché vyhľadávanie vylepšiť. Kapitola 3 sa zaoberá kontextom vyhľadávania a ako sa dá tento kontext určiť. Kapitola 4 zobrazuje niektoré existujúce podobné riešenia pre personalizáciu vyhľadávania. V kapitole 5 a 6 sme bližšie popísali návrh a implementáciu našej metódy na zisťovanie kontextu aplikácií a hľadanie aplikácie, ktorá s dopytom súvisí. Našu metódu overujeme v kapitole 7 v prostredí vyhľadávača Google. V závere sumarizujeme a naznačujeme ďalšie možné smerovanie.

Dokument pokračujeme prílohami, kde uvádzame vybranú časť technickej dokumentácie k softvérovému prototypu, ktoré je na priloženom médiu. V prílohách sa nachádza aj používateľská a inštalačná príručka. Nasleduje článok prijatý a prezentovaný na študentskej konferencii IIT.SRC 2014 a predbežná verzia príspevku pripravovaného na konferenciu ENIC 2014. Na záver uvádzame obsah priloženého média.

2 VYHĽADÁVANIE

Vyhľadávanie je neodmysliteľnou súčasťou práce s informačným priestorom. Slúži na nájdenie relevantných dokumentov, najmä vo veľkých informačných priestoroch, kde manuálne vyhľadanie často ani nie je možné. Z dôvodu informačného preťaženia sa však stáva vyhľadávanie čoraz zložitejším, a preto treba čoraz sofistikovanejšie metódy jednak na zistenie, čo sa snaží používateľ nájsť a jednak na určenie, ktoré dokumenty sú pre tento účel vhodné.

Používatelia sú zvyknutí uspokojiť sa s prvými nájdenými dokumentmi, nezávisle od ich kvality. Na vyhľadávanie si určia pomerne krátky čas, t.j. reálne pracujú len s niekoľkými prvými najlepšími výsledkami vyhľadávania. Ak však potrebnú informáciu nenájdu medzi prvými výsledkami, nie sú istí, ako majú pokračovať ďalej, aby našli očakávanú informáciu (Aula et al., 2010).

Keď používatelia nevedia nájsť výsledok, ktorý potrebujú, podľa štúdie (Aula et al., 2010) začnú:

- tráviť viac času na nájdených stránkach
- používať prirodzený jazyk pri vyhľadávaní
- meniť výrazne a nesystematicky dopyt
- rozširovať dopyt o ďalšie kľúčové slová

Používatelia teda potrebujú poradiť, ako majú ďalej pokračovať, ak nenašli očakávanú informáciu medzi prvými výsledkami vyhľadávania. Vyhľadávanie sa riadi predovšetkým dopytom, ktorý zadá používateľ. To znamená, že ak chce používateľ vhodnejšie výsledky, musí zmeniť dopyt. Používateľ teda potrebuje poradiť, ako by mohol preformulovať svoj dopyt, aby našiel požadovanú informáciu. Niekedy však pomôže, keď človek pridá k niektorému kľúčovému slovu vyhľadávačom určené metadáta, čím dá kľúčovému slovu určitú funkciu a môže usmerniť vyhľadávanie. Vyhľadávanie, v ktorom sa kľúčové slová obohacujú metadátami sa nazýva pokročilé.

2.1 Pokročilé vyhľadávanie

Vyhľadávanie len za pomoci kľúčových slov pri náročnejších úlohách nemusí postačovať. V takýchto prípadoch umožnili niektoré vyhľadávače pridať ku kľúčovým slovám ešte metadáta, ktoré špecifikujú konkrétne kľúčové slovo a dávajú mu určitú funkciu. Tabuľka 1 zobrazuje zaujímavé metadáta pre vyhľadávač Google.

""	Presná fráza, ktorá sa nemá rozdeľovať
-	Slovo, ktoré sa nemá vyhľadať
+	Slovo, ktoré sa má vyhľadať
site:	Vyhľadávanie len na určitej stránke

Tabuľka 1 Metadáta pre pokročilé vyhľadávanie

Používatelia, ktorí využívajú pokročilé vyhľadávanie majú väčšiu úspešnosť vyhľadávania a dostanú sa ku svojmu výsledku na menší počet klikov. Vyplýva to z experimentu (White et al., 2007) v ktorom porovnávali ľudí vyhľadávajúcich pomocou pokročilého vyhľadávania a len pomocou klasického vyhľadávania.

Pokročilé vyhľadávanie však využíva len okolo 10% používateľov (Jansen et al., 2000). Je to hlavne dôsledkom toho, že je to pre bežného používateľa neprirodzené. Mnoho ľudí taktiež o takejto možnosti vôbec nevie. Ľudia však prejavujú záujem o informácie ohľadom zefektívnenia vyhľadávania (Bateman et al., 2012). Radi si nechajú poradiť, ako by mohli použiť jednotlivé metadáta a postupne aj sami začnú bežne používať rozšírené vyhľadávanie, ak vidia konkrétne príklady dostatočne dlho a často.

2.2 Obohatenie dopytu

Metóda obohacovania dopytu slúži na rozšírenie používateľovho dopytu o relevantné kľúčové slová. Týmto spôsobom sa dá používateľovi pomôcť nájsť čo najlepšie výsledky jeho vyhľadávania, keďže čím lepší dopyt, tým lepšie nájdené výsledky.

Dôležité je však určiť vhodné kľúčové slová, ktorými sa obohatí dopyt. Existujú rôzne metódy, ktoré taktiež záležia na množstve a type informácií, ktoré máme o používateľovi a dopyte. Avšak v prípade, že nemáme žiadne, stále sa dajú zistiť niektoré informácie z

najvyšších nájdených výsledkov sa považuje za relevantných a informácie sa získavajú z nich, ako v (Attar et al., 1977). Potrebné informácie na obohatenie dopytu sa však zvyknú získavať pred začatím vyhľadávania. Na získanie informácií o používateľovi avšak treba mať určite práva a povolenia, od ktorých závisí rôznorodosť získaných informácií. Vyhľadávač samotný dokáže zaznamenávať predošlé dopytov používateľa (Fitzpatrick et al., 1997), no takisto dokáže získať dôležité informácie počas vyhľadávania, keď sleduje na ktoré výsledky používateľ klikol a na ktoré nie (Joachims et al., 2007). Oveľa viac informácií sa dá získať logovaním používateľa v rámci prehliadača, keď môže sledovať aj iné stránky, ktoré si používateľ prehliada a na základe týchto stránok určiť záujmy používateľa (Kramár 2014).

Obohatenie dopytu, ktorý sa deje priebežne bez používateľovej explicitnej požiadavky je dynamické. Na základe určitej implicitnej spätnej väzby, najčastejšie kliknutím na niektorý odkaz, sa obohatí dopyt a okamžite sa zmení zoznam výsledkov. Existujú rôzne spôsoby zmien výsledkov na základe dynamickej zmeny dopytu a dôležité je vhodne umiestniť nové výsledky. Jednou možnosťou je ich pridať hneď za kliknutý odkaz a vizuálne ich odsadiť, ako (Kim et al., 2013). Avšak takto to pre používateľa vyzerá, že nové výsledky súvisia len s kliknutým odkazom a keďže používateľ nenašiel odkaz na danej stránke, môže tieto odporúčania preskočiť bez čítania. Takisto medzi výsledkom, za ktorý pribudli odkazy a tým ďalším, vznikne iné vyhľadávanie bez možnosti ďalšieho dynamického obohatenia dopytu

Inou možnosťou je zamiešať nové výsledky medzi aktuálne a zoznam výsledkov tak neustále meniť, respektíve obmieňať. Problém však nastáva, keď používateľ znova zadá rovnaký dopyt a musí sa zobraziť pôvodný počet výsledkov. Ľudia si však nepamätajú celý zoznam výsledkov vyhľadávania. Pamätajú si len kľúčové položky v danom zozname (Teevan, 2008). Sú to tie výsledky, ktoré im výrazne pomohli, navyše aj prvý výsledok. Vďaka týmto kľúčovým položkám je tak ich neskoršie vyhľadávanie rýchlejšie. Ak však zmiznú, alebo sa presunú, používatelia považujú tento zoznam za úplne nový. Preto ak nahradíme len neužitočné výsledky vyhľadania za užitočné, získame nový zoznam, v ktorom môže používateľa vyhľadávať rovnako jednoducho a navyše má lepšie výsledky.

2.3 Personalizované vyhľadávanie

Personalizácia je proces prispôsobovania sa obsahu a štruktúry systému špecifickým a individuálnym potrebám každého používateľa (Eirinaki et al., 2003). Dôležité je preto poznať používateľa a jeho správanie. Vo veľkej miere sa to dá dosiahnuť pozorovaním používateľa v prostredí počítača a analýze získaných dát. Pozorovať však môžeme rôzne aspekty práce používateľa s počítačom, či už ako jednotlivca, alebo člena určitej skupiny.

2.3.1 Personalizácia na základe profilu používateľa

Pomocou tejto metódy sa zohľadňuje používateľ ako indivíduum a berú sa do úvahy jeho osobné vlastnosti a skúsenosti. Cieľom je vytvoriť profil používateľa, ktorý bude zahŕňať jeho osobné preferencie.

2.3.1.1. Personalizácia na základe histórie vyhľadávania

V tejto metóde sa zohľadňuje používateľova história vyhľadávania a stránky, na ktoré klikal. Ak zadá neskôr používateľ rovnaký, alebo podobný dopyt, zobrazia sa mu už len tie výsledky, ktoré ho zaujali. Touto metódou sa dá značne používateľovi pomôcť v prípade, že používateľ používa vyhľadávanie napríklad na presmerovanie sa na určitú stránku, pričom vždy zadá rovnaký dopyt a očakáva zakaždým tú istú stránku bez ohľadu na akékoľvek okolnosti. Táto metóda je však nefunkčná, keď používateľ zadá dopyt 1. krát.

2.3.1.2. Personalizácia na základe používateľových záujmov

Na základe tohto druhu personalizácie sa zisťujú používateľove záujmy. Záujmy sa dajú zistiť z predošlých vyhľadávaní a interakcie, keďže používateľ vyhľadáva to, čo ho zaujíma. Do úvahy sa dajú brať:

a) dlhodobé záujmy: všetky získané záujmy zo všetkých sedení
b) krátkodobé záujmy: záujmy, ktoré používateľa získal pri aktuálnom sedení
c) kombinácia krátkodobých a dlhodobých záujmov

Touto metódou môžeme používateľovi pomôcť v prípadoch, keď často hľadá informácie ohľadom svojich koníčkoch, ktoré sa až tak často nemenia. V takomto prípade vyhľadávanie zúžime len na tie oblasti, ktoré súvisia s koníčkom, ktorý aktuálne vyhľadáva.

2.3.2 Kolaboratívna personalizácia

Ľudia s podobnými záujmami klikajú na podobné výsledky. Toto je predpoklad kolaboratívnej personalizácie. Základom je správne rozdelenie do skupín podľa dlhodobých záujmov. Keď niektorý používateľ zadá dopyt, výsledky sa zohľadnia podľa klikaní členov skupín. O relevancii pritom rozhoduje obľúbenosť daného výsledku v skupine. Tento prístup bol použitý v (Sun et al., 2005), kde používatelia boli rozdeľovaní do skupín podľa korelácie ich dopytov a stránok, na ktoré klikali. Presnejšie ich vieme rozdeliť ak ich pozorujeme v širšom prostredí ako len počas vyhľadávania, napríklad neustále pri používaní prehliadača. Týmto ich môžem zaradiť do skupín aj na základe bežných aktivít a určiť tak presnejšie skupinu, do ktorej používateľ patrí.

2.3.3 Porovnanie

Ako tieto rôzne prístupy zlepšujú vyhľadávanie sa zaoberali (Dou et al., 2007) vo svojom experimente, v ktorom skúmali logy 10 000 používateľov. Zistili, že všetky metódy výrazne vylepšujú vyhľadávanie na výsledkoch s malou variáciou klikaných odkazoch a naopak, kde používatelia klikali na rôzne výsledky, bolo personalizované vyhľadávanie nevhodnejšie. Najlepšie výsledky dosiahla personalizácia na základe histórie používateľa.

2.4 Zistenie kvality vyhľadávača

Aby sme vedeli porovnať, ktorý vyhľadávač nám dáva kvalitnejšie výsledky, je treba skúmať, kedy používateľ úspešne ukončil vyhľadávanie, respektíve neúspešne a koľko úsilia na to musel vynaložiť. Zistiť, či používateľ úspešne alebo neúspešne ukončil vyhľadávanie a či vôbec ukončil vyhľadávanie nie je vôbec triviálne. Ukončenie vyhľadávania by bolo možné zistiť časovým limitom, avšak to nie je dostačujúce (Jones et al., 2008). Ak sa totiž stanoví malý čas, súvislé vyhľadávanie sa rozdelí na niekoľko častí aj v prípade, že používateľ si chce nájdenú informáciu overiť. Overenie ľahko zaberie dlhší čas ako sa stanovil na oddelenie vyhľadávaní a ak je daná informácia nedostačujúca a používateľ znova vyhľadáva rovnakú vec, toto vyhľadávanie sa už pokladá za nové. Ak sa naopak určí dlhší čas, po ktorom sa považuje vyhľadávanie za nové, používateľ môže vyhľadávať už niečo iné, no neuplynul stanovený časový limit a tak sa považuje toto nové vyhľadávanie len za pokračovanie predošlého. Najvyššiu presnosť 71.2% dosiahol časový limit 5 minút (Jones et al., 2008), avšak v (Jiang et al., 2011) dosiahol najvyššiu presnosť zas časový limit 25-30 minút.

Pri skúmaní, koľko námahy na dané vyhľadávanie používateľ vynaložil viaceré výskumy berú ako metriky len poradie výsledkov, na ktoré používateľ klikol. Na zistenie kvality personalizovaného vyhľadávania sa dá použiť metóda AverageRank (Dou et al., 2007; Speretta et al., n.d.; Qiu et al., 2006), ktorá pre konkrétne vyhľadávanie *s* určí, aké je priemerné poradie kliknutých výsledkov. Čím viac kliknutých výsledkov bolo na vrchných pozíciách, tým menšia hodnota AverageRank bude a tým bolo vyhľadávanie lepšie.

$$AvgRank_s = \frac{1}{|P_s|} \sum_{p \in P_s} R(p)$$

kde P_s sú všetky kliknuté výsledky jedného dopytu a R(p) je poradie daného výsledku.

Táto metóda však neberie do úvahy, aké relevantné dokumenty sú tie, na ktoré používateľ klikol. Presnejšiu metódu poskytuje Markov model pravdepodobnosti, ktorý zohľadňuje aj čas medzi kliknutím na ďalší výsledok. Avšak tieto modely nezohľadňujú prácu používateľa s kliknutým výsledkom. Ak napríklad používateľ dokument hneď zavrie, znamená to že dokument je pravdepodobne rovnako nevhodný, ako keby naň ani neklikol.

3 KONTEXT VYHĽADÁVANIA

Existuje niekoľko definícií kontextu. Jednotlivé prístupy k definovaniu tohto pojmu sú často ovplyvnené účelom použitia kontextu pri práci s informáciami. Snáď najčastejším pohľadom na kontext je definícia uvedená v (Dey et al., 1999): *„kontext je akákoľvek informácia, ktorá môže byť použitá na charakterizovanie okolnosti nejakej entity. Entita je osoba, miesto, alebo objekt, ktorý je považovaný za relevantný v závislosti k interakcii používateľa s aplikáciou, vrátane samotného používateľa a aplikácie.“*

Podľa definície sa dá usúdiť, že kontext je akákoľvek informácia, ktorá ovplyvňuje používateľa pri vykonávaní určitej úlohy v aplikácií. Získavanie týchto informácií je však náročné, pretože jednak sa nedá vo všeobecnosti určiť, ktoré informácie ovplyvňujú používateľa a jednak je problém vôbec so zaznamenávaním týchto informácií. Tieto informácie sa získavajú buď explicitne alebo implicitne. Explicitne znamená, že používateľ poskytuje tieto informácie priamo. Získavanie kontextu explicitne však môže byť problematické, keďže to zaťažuje používateľa a používateľ nemusí správne odhadnúť, čo všetko sa v momentálnej úlohe považuje za kontext. Inou možnosťou je implicitná metóda, pri ktorej používateľ nedáva priamo túto informáciu, ale získavame ju odvodením z jeho správania.

Keďže je všetkých informácií o kontexte používateľa veľa, bolo by vhodné ich kategorizovať a určiť, ktoré sú najdôležitejšie pre určenie kontextu vyhľadávania. Za vhodné kategórie navrhol (Ryan et al., 1998) polohu, prostredie, identitu a čas. Z iného hľadiska považujú (Schilit et al., 1994) za dôležité aspekty toho, kde som, s kým som a čo je naokolo. Táto definícia však zohľadňuje iba identitu a polohu. Za najvhodnejšie kategórie zvolil (Dey et al., 1999) polohu, identitu, aktivitu a čas.

3.1 Kontext aktivity

Z pohľadu aktivity sa dá kontext chápať ako interakčný problém, nie reprezentačný. To znamená, že kontext aktivity je udávaný dynamicky a aktívne vytváraný, nedá sa vopred deklarovať. Predchádzajúca sociálna interakcia, tak isto ako predošlé využívanie mediálnych prostriedkov, nástrojov a prístrojov ktoré každodenne používame ovplyvňujú naše aktuálne aktivity (Schilit et al., 1994). V prostredí počítača pod aktivitou, ktorá

ovplyvňuje kontext vyhľadávanie, môžeme rozumieť napríklad používanie rôznych aplikácií, prehliadanie rôznych dokumentov a aj vyhľadávanie samotné sa stáva kontextom aktivity.

Predpokladajú sa tieto vzťahy medzi kontextom a aktivitou:

- každá aktivita, napríklad písanie dokumentu, vyhľadávanie, má kontext. Táto aktivita dynamicky podnecuje vytváranie ďalšieho kontextu
- kontext je to, čo dáva zmysel aktivite. Dôsledkom toho je, že kontext bez aktivity neexistuje
- každá aktivita môže byť kontextom inej aktivite. Kontext je teda tvorený zjednotením kontextových informácií vznikajúcich. To znamená, že písanie dokumentu môže byť kontextom aj inej aktivite, ako napríklad vytváraniu obrázku v grafickom editore a kontext aktivity v grafickom editore by bol teda zjednotením kontextu aktivity v grafickom editore a písania dokumentu

3.2 Zohľadnenie kontextu

Zohľadnenie kontextu v informačnom systéme je čoraz populárnejšie. Systém sa pri tom však vo väčšine prípadov nesnaží odhadnúť alebo získať kontext používateľa, ale vývojár systému odhaduje, v akom kontexte sa môže používateľ na danej stránke nachádzať. Ak je to napríklad internetový obchod zameraný na strojárske potreby, vývojár predpokladá, že používateľ pracuje v strojárskom priemysle a má v pláne si niečo kúpiť. Webový portál zaoberajúci sa maturitnými otázkami zas predpokladá, že používateľ je študent, ktorý sa pripravuje na maturity. To vývojárovi pomôže upraviť stránku tak, aby uľahčil používateľovi navigáciu na stránke a nasmerovať ho priamo k jeho predpokladanému cieľu.

Informačné systémy zohľadňujúce kontext používateľa typicky vykonávajú tieto úlohy (Polzonetti, 2010):

- oboznámenie sa: zbierka relevantných dát umožňujúcich objasniť kontext
- zdôvodňovanie: dedukcia zo surových dát na abstraktné informácie
- učenie a predpovedanie na základe akcií a kontextových informácií z minulosti

- reprezentácia: modelovanie kontextovej informácie do strojovo-čitateľnej a zrozumiteľnej formy
- manažment a rozšírenie kontextovej informácie
- aktivácia: zmena správania na základe dostupnej kontextovej informácie

3.3 Získavanie informácií od používateľa

Vzhľadom k tomu, že sa snažíme vylepšiť výsledky vyhľadávania na základe jeho kontextu aktivity, je veľmi dôležité zvoliť vhodný prístup ku získaniu týchto informácií. Tieto informácie nevieme bez neustálej spätnej väzby používateľa zistiť, pretože aktivita na počítači sa mení a teda aj jej kontext. K získaniu kontextu aktivity je preto nevyhnutná spätná väzba používateľa. Všeobecne známe sú 2 prístupy k získaniu spätnej väzby: implicitná a explicitná

3.3.1 Explicitná spätná väzba

Spätná väzba, ktorú poskytuje používateľ úmyselne sa nazýva explicitná spätná väzba. Keďže ju poskytuje používateľ osobne, z vlastného rozumného uváženia a zo svojej vlastnej vôle, dá sa predpokladať, že presnosť je veľmi vysoká. Dávanie tejto spätnej väzby však zaťažuje používateľa a preto ju poskytne len zriedka. Navyše hrozí, že ho to natoľko odradí, že používateľ prestane reagovať. Hrozí aj, že používateľ prestane venovať explicitnej spätnej väzby dostatočnú pozornosť a poskytne nepresné informácie. Je preto dôležité pýtať si spätnú väzbu vhodne.

Používatelia sú ochotní poskytovať spätnú väzbu, pokiaľ im to pomôže pri vyhľadávaní. Vyplýva to z experimentu (Lagun et al., 2013), v ktorom zisťovali, ako sa ľudia stotožnia s možnosťou explicitnej spätnej väzby pri vyhľadávaní a ako im pomôže. Pri vyhľadávaní sa im zobrazovala možnosť, či si prajú definovať mesto, v ktorom chcú vyhľadať zadanú úlohu. Výsledne vyšlo, že vyhľadávanie so spätnou väzbou má väčšiu úspešnosť aj efektivitu, avšak štatisticky málo významnú. Keď sa však respondentov pýtali, či by chceli mať zahrnutú túto možnosť vo svojom prehliadači, až 75% respondentov odpovedalo pozitívne. Tí, čo odpovedali negatívne dodali, že vyhľadávač by mal danú informáciu vedieť. Takisto až v 78% prípadoch účastníci experimentu poskytli spätnú väzbu, keď sa im naskytla možnosť a mali indíciu, že to pomôže pri vyhľadávaní.

Spätnú väzbu môžeme vo všeobecnosti rozdeliť navyše aj na pozitívnu a negatívnu. Pozitívnu spätnú väzbu dáva používateľ, keď informácia vystihuje problematiku a negatívnu spätnú väzbu, keď informácia zavádza a z používateľovho hľadiska nesúvisí s problematikou. Vďaka pozitívnej spätnej väzbe sa pri vyhľadávaní dá lepšie vymedziť, čo presne používateľ myslel a cez negatívnu spätnú väzbu vylúčiť tie oblasti, ktoré síce môžu súvisieť s dopytom, no pre daný kontext používateľa nemajú žiadny zmysel.

Vo svojom experimente (Peska et al., 2013) skúmali dôsledok negatívnej spätnej väzby pri odporúčaní v elektronickom obchode len na základe negatívnej explicitnej spätnej väzby. Zistili, že odporúčanie na základe len tejto negatívnej spätnej väzby bolo dosť nepresné, presnejšie dokonca bolo aj odporúčanie len na základe používateľových preferencií. Avšak v prípade, že používateľ dostal 5 alebo 10 vhodných výsledkov sa táto metóda osvedčila a dosiahla oveľa lepšie výsledky ako iné metódy. Z tohto sa dá usúdiť, že negatívna spätná väzbe je dôležitá pri utrieďovaní vhodných objektov.

3.3.2 Implicitná spätná väzba

Spätná väzbu, ktorú používateľ poskytuje bez toho, aby vnímal, že ju poskytuje sa nazýva implicitná. Predpokladá sa, že používateľ chce vidieť len dokumenty, ktoré ho zaujímajú a minimalizovať čas strávený na tých, ktoré idú mimo jeho oblasť záujmu (Pirolli et al., 1995). Z tohto dôvodu sa dá len z pozorovania používateľa usúdiť, ktoré dokumenty sú pre používateľa relevantné a ktoré nie. Hodnovernosť takto získaných informácií je nízka, pretože počítač dedukuje z momentálnej aktivity, čo používateľove akcie znamenajú. Keďže získavanie týchto dát nezaťažuje používateľa, nie je problém získať veľmi veľa dát. Problematické je však určiť, ktoré dáta sa majú zbierať pre čo najpresnejší úsudok a ako reprezentovať používateľovu spätnú väzbu a jeho akcie.

Veľmi presnou metódou na získanie implicitnej spätnej väzby ohľadom kvality dokumentu je zariadenie na snímanie pohľadu, Eyetracking. Zaznamenáva časti dokumentu, na ktoré sa používateľ pozerá a podľa tohto je možné jednoducho získať tie úseky dokumentu, ktoré používateľa zaujali. Toto zariadenie je však málo dostupné a bežný používateľ ho nemá k dispozícií.

Používateľ kliká na dokumenty, ktoré ho zaujmú, či už svojím nadpisom, alebo krátkym popisom. Z tohto predpokladu vychádzajú vo svojich experimentoch (Joachims et al.,

2007; Jiang et al., 2011). Dôležitou implicitnou spätnou väzbou je preto už len kliknutie, resp. otvorenie dokumentu. Toto však ešte neznamená, že je dokument relevantný, pretože môže kliknúť z omylu, alebo dokument neobsahuje očakávané informácie. Ďalším faktorom preto je aj pozorovanie, ako dlho na danom dokumente zotrval. Predpokladá sa, že čím dlhšie, tým je dokument relevantnejší (Leung & Lee 2010). Implicitnými faktormi relevancie dokumentu sú takisto kopírovanie a označovanie textu, pričom používateľ kopíruje tie časti textu, ktoré ho zaujali a chystá sa použiť aj v inej aplikácií.

3.4 Existujúce kontext zohľadňujúce vyhľadávania

Moderné webové vyhľadávače sa už snažia zohľadniť kontext používateľa. Ich schopnosť získavania kontextu používateľa je však značne obmedzená, lebo môžu zaznamenávať aktivitu a záujmy používateľa len v rámci svojej domény. Z tohto dôvodu sú najdôležitejšími a mnohokrát jedinými údajmi, ktoré sa zaznamenávajú, údaje o dokumentoch, na ktoré používateľ klikol a dopyt, ktorý používateľ zadal. Vyhľadávania, ktoré zohľadňujú kontext sa delia na 2 skupiny a to na vyhľadávače samotné a na projekty, ktoré upravujú vyhľadávanie v niektorom vyhľadávači na základe ich metód personalizácie vyhľadávania.

Podrobným skúmaním implicitnej spätnej väzby sa venovali vo svojej práci (Joachims et al., 2007). Skúmali ako vplýva rôzne vyhodnocovanie implicitnej spätnej väzby na určenie aktuálneho kontextu za účelom obohatenia dopytu. Zamerali sa na ohodnotenie relevantnosti tých výsledkov, na ktoré používateľ klikol, ale ja na tie, na ktoré používateľ neklikol.

Experiment spočíval v splnení určitej úlohy, či už nájdenie určitej špecifickej stránky, alebo špecifickej informácie. Následne požiadali porotcov experimentu aby zoradili výsledky od najlepšieho po najhorší. Skúmali, či relevancia dokumentov reprezentovaná implicitnou väzbu experimentátorov na základe niekoľkých stratégií zodpovedá relevancii určenej expertmi.

Určili si niekoľko stratégií, ktorými ohodnocovali relevanciu dokumentu, ak napríklad používateľ klikne na niektorý výsledok vyhľadávania a na vyšší neklikol, ten nižší má vyššiu relevanciu. Pomocou nich sa dá určiť pomerne presne relevancia daného odkazu. Avšak odhalili aj šum v podobe poradia výsledkov, vďaka čomu dokázali určiť len

relatívnu relevanciu, nie absolútnu. To znamená, že dokázali určiť relevanciu len vzhľadom k poradiu daných výsledkov.

Kolektív autorov (Jiang et al., 2011) vyvinuli personalizačný rámec, ktorý sa snažil zistiť používateľov kontext a na základe neho usporiadať výsledky, aby na najvyšších priečkach boli tie, ktoré súvisia s používateľovým kontextom. V tomto experimente analyzovali logy z vyhľadávacieho nástroja a pokúšali sa ich rozdeliť na jednotlivé sedenia. Pomocou časového limitu a SEPR podobnosti sa im podarilo dosiahnuť vysokú presnosť (99.4%) avšak nízku integritu (40.28%), čo znamená, že odhalili takmer vždy ukončenie sedenia, avšak odhaľovali ukončenie sedenia aj tam, kde nemali, kvôli prísnym pravidlám SERP podobnosti. SERP podobnosť je kosínusová podobnosť výsledkov 2 za sebou idúcich dopytov a keď je menšia ako zvolený prach, považuje sa za nové sedenie.

Najvyššia presnosť (96,89%) a integrita (78,36%) v odhaľovaní sedení bola dosiahnutá kombináciou časového limitu, SERP podobnosti a zistenia novej reformulácie. Reformulácia je zmena dopytu podľa niektorého z ich pravidiel, ako napríklad pridanie slova, rozšírenie skratky, zmena poradia slov a iné. Avšak identifikáciu reformulácie dopytu robili ľudskí porotcovia.

Používateľov kontext síce získavali podľa rôznych zaužívaných metód (Joachims-C, mJoachims-C) z predošlých podobných dopytoch, avšak dokázali s vysokou úspešnosťou rozdeľovať rôzne sedenia a zistili, že ak použijú rôzne stratégie pre rôzne typy reformulácií, dostanú lepšie výsledky.

Skúmaniu vplyvu pozitívnej a negatívnej spätnej väzby sa podrobnejšie venovali (Leung et al., 2010). Analyzovali a zohľadňovali negatívnu spätnú väzbu. Navyše používateľov profil budovali na základe konceptov a nie dokumentov, t.j. používateľov profil obsahoval tematické oblasti a nie dokumenty. Používateľov profil sa skladal z pozitívnych a negatívnych konceptov. Zohľadnením negatívnej spätnej väzby dokázali zvýšiť úspešnosť personalizovaného vyhľadávania. Experiment vyzeral podobne ako v prípade (Joachims et al., 2007), spočíval teda v analyzovaní klikov a porovnávaní s ľudskými porotcami.

3.5 Diskusia

Určovanie kontextu vyhľadávania je náročné, pretože existuje mnoho faktorov, ktoré naň vplývajú a takisto sa zložito zaznamenávajú. Zvolili sme si určovanie kontextu vyhľadávania na základe kontextu aktivity. Kontext aktivity je však veľmi dynamický a rýchlo sa mení. Určovať ho explicitnou spätnou väzbou preto nie je realizovateľné. Dá sa však odhadnúť pomocou implicitnej spätnej väzby, teda sledovaním používateľových aktivít.

Vyššie spomínané existujúce riešenia skúmali a vyhodnocovali používateľovu aktivitu hlavne pri vyhľadávaní dopytov, kedy na základe jeho klikov dedukovali relevanciu dopytov a jeho aktuálny cieľ. Domnievame sa, že kontext vyhľadávania sa dá zistiť sledovaním používateľovej aktivity už pred začatím vyhľadávania v rámci všetkých aplikácií a na základe zaznamenanej aktivity indukovať používateľov cieľ.

4 NÁVRH METÓDY OBOHACOVANIA DOPYTU

V predchádzajúcej kapitole sme analyzovali už existujúce riešenia. Existujúce riešenia zohľadňujú používateľa len v priestore prehliadača. Domnievame sa však, že aj aktivita mimo prehliadača, resp. činnosti, ktoré používateľ vykonáva s jej kontextom (objekty, s ktorými pracoval) nám poskytne cenné vstupy pre presnejšie odhalenie používateľovho kontextu aktivít, keďže používateľ používa aj iné aplikácie pri práci s počítačom. Pre používateľa, ktorý pracuje momentálne s desktopovými aplikáciami je preto vhodnejšie získavať kontext aj z tých aplikácií, ktoré používa. Nami navrhnutá metóda pre vyhľadávanie s využitím kontextu aktivity z tohto dôvodu skúma používateľa aj mimo prostredia vyhľadávača, dokonca mimo priestoru prehľadávača. Snaží sa získať čo najširší používateľov kontext zo všetkých aplikácií a na základe kontextu aplikácií obohatiť dopyt.

Rôzne aplikácie však slúžia na rôzne účely, preto má používateľ v rôznej aplikácii rôzny kontext aktivity. Z tohto dôvodu sa nedá obohatiť dopyt o kontext aktivity všetkých aplikácií, ale len tej, z kontextu ktorej používateľ pri vyhľadávaní vychádzal. Dôležité teda je nájsť súvislosť medzi kontextom aktivity niektorej z aplikácií a dopytom, ktorý používateľ vyhľadáva. Ak sa podarí nájsť, tak sme získali širší kontext daného dopytu, pretože predpokladáme, že kontext daného dopytu by mal byť rovnaký, resp. podobný ako kontext aplikácie, z ktorej vychádzal. Vzhľadom k tomu, že sa zaoberáme skúmaním len kontextu aktivity, kontext aplikácie v našej práci je reprezentovaný kontextom aktivity v danej aplikácií. Práve kontext aplikácie, ktorý je rovnaký alebo podobný kontextu vyhľadávania, vyhľadávač potrebuje vedieť, aby vedel určiť správny význam viacvýznamového dopytu, prípadne aby usmernil dopyt niektorým smerom. Ak napríklad používateľ vyhľadáva názov filmu, môžeme jeho výsledky usmerniť smerom k nájdeniu titulkov k danému filmu, pokiaľ tomu odpovedá aktuálny kontext daného prehrávača filmov.

Nie je však možné odovzdať celý nami získaný kontext vyhľadávaču z dôvodu prílišného množstva informácií. Dôležité je preto vybrať zopár kľúčových slov, ktoré budú tento kontext vyjadrovať čo najpresnejšie a ktoré môžu vyhľadávaču pomôcť nájsť požadovaný výsledok. Nami navrhovaná metóda na obohatenie dopytu teda pozostáva z týchto krokov:

1. zachytenie vyhľadávania
2. získanie kontextu aktivity používateľa
3. nájdenie prepojenia medzi kontextom aplikácie a dopytom
4. obohatenie dopytu používateľa

4.1 Získavanie kontextu aktivity používateľa

Vzhľadom k tomu, že získavanie kontextu aktivity používateľa explicitnou spätnou väzbou je nemožné kvôli nutnosti neustáleho sa dopytovania používateľa, kontext aktivity sa musí získavať implicitnou spätnou väzbou, t.j. z aktivity používateľa v počítači.

Informácie získané implicitnou spätnou väzbou však obsahujú aj balast, ktorý nám neposkytuje žiadne cenné informácie, preto ho treba ešte pred prevedením na aktuálny kontext aplikácie upraviť do vhodnej formy. Nami navrhnutá metóda preto spočíva v nasledovných krokoch:

1. zachytenie aktivity používateľa
2. spracovanie záznamu aktivity

Kroky metódy nie sú výpočtovo náročné, preto sa môžu vykonávať za behu.

4.1.1 Zachytenie aktivity používateľa

V rámci všetkých aplikácií je zachytenie aktivity používateľa zložité, pretože každá aplikácia má svoju špecifickú architektúru a s operačným systémom spolupracuje len na vrchnej vrstve. To znamená, že nie je možné napojiť sa na každú aplikáciu a zisťovať používateľove akcie v aplikácií. Implicitnú spätnú väzbu teda odchytávame na úrovni operačného systému. Operačný systém poskytuje aplikačné rozhranie, vďaka ktorému je možné odchytávať viaceré udalosti, ktorými aplikácie komunikujú s operačným systémom, napr. zmena aktívnej aplikácie. Odchytávanie týchto udalostí nám poskytuje informácie, vďaka ktoré môžeme určiť kontext aplikácie.

Na vrstve operačného systému je možné napojiť sa priamo na aplikácie založené na windowsovom rozhraní. Toto nám umožňuje získať kontext aj z textových polí danej aplikácie. Získame tým informácie o tom, čo práve používateľ číta a píše vo všetkých aplikáciách. Funguje to však len pre aplikácie postavené na windowsovom rozhraní

a priame napájanie do aplikácií je nebezpečné, nečakané výnimky môžu spôsobiť pád aplikácie, preto sme túto možnosť zamietli.

Internetový prehliadač je však zvláštnym typom aplikácie. Zobrazuje webové aplikácie, ktoré ponúkajú používateľovi možnosti aj desktopových aplikácií, preto sa účel a kontext tejto aplikácie stáva účelom a kontextom webovej aplikácie, ktorú používateľ momentálne používa. Aká to je však zistiť na vrstve operačného systému nevieme. Táto informácia je však veľmi cenná pre zistenie kontextu používateľa, keďže internetové aplikácie sa stávajú čoraz populárnejšími a lákavejšími. Rozšírenie do internetového prehliadača je preto nutnosťou. Rozšírenie je schopné napojiť sa priamo na internetový prehliadač pomocou skriptu a teda odchytávať viaceré udalosti na úrovni internetového prehliadača. Dokáže dokonca do každej webovej aplikácie vložiť vlastný skript, čím vie získať ľubovoľnú informáciu z danej stránky a ľubovoľnú udalosť.

Používateľ zvykne vyhľadávať v digitálnych knižniciach práve vtedy, keď píše odborný článok. Keďže Microsoft Office Word je najpoužívanejšia aplikácia pre písanie dokumentov, rozhodli sme sa získavať širší kontext práve z tejto aplikácie. Microsoft Office umožňuje vyrobiť do svojich produktov add-in, ktoré pridá produktu potrebnú funkcionalitu. Pre získanie širšieho kontextu z Office Word aplikácie je potrebný add-in, ktorý sa po nainštalovaní dokáže napojiť na ľubovoľný dokument upravovaný touto aplikáciou. Vďaka tomu vieme získať dostatok informácií na určenie aktuálneho kontextu. Keďže dokumenty bývajú väčšinou rozsiahle, predpokladáme, že práve upravovaný text je centrom používateľovej pozornosti a aktuálneho záujmu, preto sa sústredíme práve naň.

4.1.2 Spracovanie záznamu aktivity

Získaný záznam aktivity je potrebné preformulovať na kľúčové slová, ktoré budú vyjadrovať danú aktivitu. Kľúčové slová sú pre počítač ľahšie spracovateľné a vie s nimi jednoduchšie manipulovať ako keby bola aktivita vyjadrená prirodzeným jazykom.

Prirodzený jazyk obsahuje informácie, ktoré sú pre zistenie kontextu nevhodné, lebo samé o sebe neobsahujú žiadnu informáciu. Sú to takzvané stop slová. Týchto slov však nie je veľa a tak ich odstránenie je skoro okamžité.

Prirodzený jazyk je špecifický aj svojím tvaroslovím. Slová, ktoré svojou podstatou vyjadrujú rovnakú vec majú rôzny tvar pre iné gramatické kategórie, napríklad slovo žena

môže nadobudnúť rôzne tvary, napr. ženy, žien, ženami a podobne. Z kontextového hľadiska vyjadrujú tieto slová rovnaký význam, avšak algoritmus to nevie určiť. Dôležité je preto previesť tieto slová na ich základný tvar, v ktorom bude algoritmus vedieť, že sú tieto slová rovnaké. K tomuto cieľu slúži stematizácia slova.

Spracovanie prirodzeného jazyka je však závislé na jazyku. Každý jazyk má svoje vlastné stop slová a každý jazyk stematizuje slová inak. Preto pokiaľ nevieme určiť jazyk textu, nemôžeme upravovať prirodzený jazyk týmito metódami. Na malom množstve textu je však zložité určiť jazyk, pretože viaceré jazyky majú rovnaké slová. Pre zjednodušenie a získanie dôveryhodnejších informácií sme zúžili množinu jazykov na dva: anglický a slovenský. Pre každý jazyk sme následne ignorovali ich stop slová a zvolili vhodnú metódu na stematizovanie slov.

Avšak nie všetky získané informácie sú kontextom domény. Pri niektorých typoch akcií sa získava informácia, ktorá síce slúži danej aplikácií k určitému účelu, avšak pre nás nemá žiadnu hodnotu. Identifikujeme ich na základe toho, že sa prenášajú pri každom zachytení udalosti rovnakého typu, napr. aplikácie zvyknú do názvov svojich okien vkladať aj názov aplikácie. Keďže sa však vyskytujú zakaždým pri rovnakom type udalostí v nezmenenej forme, nie je ich ťažké identifikovať a odfiltrovať.

4.2 Nájdenie prepojenia medzi aplikáciou a dopytom

Aby sme mohli obohatiť dopyt používateľa o relevantné kľúčové slová získané z kontextu niektorej aplikácie, ktorú nedávno používal, je potrebné určiť tú správnu, keďže rôzne aplikácie majú rôzny kontext.

Hlavným cieľom je nájsť prepojenie medzi dopytom a kontextom niektorej aplikácie. Dopyt používateľa je však často zadávaný bez diakritiky, pretože väčšina vyhľadávačov diakritiku ignoruje a automaticky nájde aj výsledky, pri ktorých je diakritika odlišná. Avšak keď spracovávame prirodzený jazyk a potrebujeme ho upravovať, potrebujeme vedieť správny tvar slova. Keďže zistenie správnej diakritiky slova nie je účelom našej práce, navrhli sme 2 jednoduché metódy na zistenie správnej diakritiky.

Prvá metóda berie do úvahy obrovskú dátovú sadu rôznych viet, získaných napríklad spracovaním kníh, článkov na Wikipédií alebo rôznych iných zdrojov. Tieto vety rozbijeme na jednotlivé slová a porovnávame tieto slová s dopytom, aby keď sa nájde také,

ktoré je bez diakritiky rovnaké ako dopyt požívateľa, mohol priradiť diakritiku. Ak existujú rôzne slová, ktoré bez diakritiky dávajú to isté slovo, vyberie sa to, ktoré sa vyskytovalo v dátovej sade najčastejšie. Touto metódou sa však môže niektorým slovám priradiť chybná diakritika v prípade ak používateľ myslel na slovo, ktoré sa v dátovej sade nevyskytuje najčastejšie.

Inou metódou môžeme využiť tie výsledky, ktoré vyhľadávač našiel a použiť ich na zistenie správnej diakritiky. Vyhľadávač síce vyhľadáva tak, že ignoruje diakritiku, avšak nájdené výsledky sú zobrazené so správnou diakritikou. Preto ak porovnáme náš dopyt s nájdenými výsledkami, mohli by sme vo výsledkoch nájsť slová dopytu so správnou diakritikou. Úspešnosť pridania správnej diakritiky je vtedy rovná úspešnosti, s akou vyhľadávač priraďuje diakritiku.

So správne zadaným dopytom môžeme presnejšie zisťovať, ktorá aplikácia súvisí s dopytom. Dopyt musíme predspracovať rovnako ako sa spracúva aktivita, pretože je rovnako zadávaný v prirodzenom jazyku. Vďaka správnej diakritike je určenie jazyka presnejšie. Pridaná diakritika umožňuje aj stemovanie kontextu, vďaka ktorému vieme aj kľúčové slová dopytu previesť na základný tvar. Z dopytu je rovnako potrebné odstrániť stop slová, ktoré aj samotný vyhľadávač ignoruje. Až s takýmto predspracovaným dopytom začíname hľadať spojitosť k niektorej z aplikácií.

Navrhli sme rozhodovaciu metódu podľa ktorej budeme určovať kontext aplikácie, ktorá súvisí s dopytom. Zameriavame sa na aktuálny kontext aktivity, preto budeme do úvahy brať len kontext určitého posledného krátkeho obdobia a v ňom hľadať súvislosť s dopytom.

Prvou metódou, ktorou sa pokúšame nájsť spojitosť medzi dopytom a kontextom aplikácie je syntaktická zhoda. To znamená, že hľadáme prekryv kľúčových slov dopytu a kľúčových slov kontextu aplikácie. Kľúčové slová sú už predspracované a upravené na základný tvar, čo nám túto akciu výrazne uľahčuje. Vzhľadom k tomu, že stemované kľúčové slová zahŕňajú v sebe aj určitú nepresnosť, väčšiu váhu prideľujeme prekryvu v prirodzenom jazyku, t.j. v tvare, v ktorom boli zadané alebo získané.

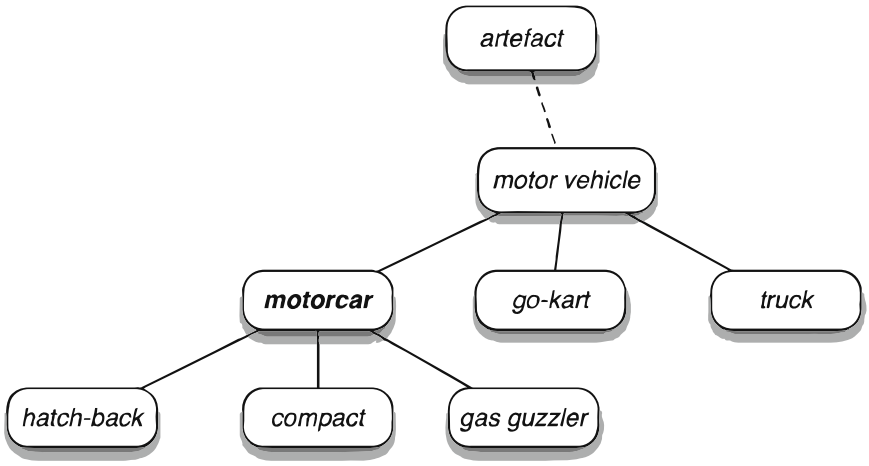
Práve z kontextu správnej aplikácie sa snažíme zistiť špecifický zmysel používateľovho dopytu, preto ak určíme nesprávnu aplikáciu, môžeme dopyt rozšíriť o nesprávne kľúčové

slová, čím by sme používateľovi mohli vyhľadávanie ešte viac znepresniť. Z tohto dôvodu určujeme, že kontext niektorej aplikácie súvisí s dopytom len v prípade, že sa nájde prekryv kľúčových slov a kontextu tejto aplikácie. Ak sa nájde prekryv len v jedinej aplikácií, môžeme predpokladať s vysokou pravdepodobnosťou, že dopyt súvisí práve s touto aplikáciou. V inom prípade aplikujeme ešte ďalšiu vyhodnocovaciu metódu.

Prirodzený jazyk sa vyznačuje aj tým, že pre rovnaké slovo má rôzne tvary, synonymá. Slová, ktoré majú rovnaký význam, avšak sa inakšie píšu. Z tohto dôvodu nám syntaktická analýza neodhalí, že ide o rovnaké alebo veľmi podobné slovo, ak je vyhľadávané v niektorom synonymickom tvare. Aby sme však zistili, že ide o rovnaké alebo veľmi podobné slovo, použijeme sémantickú analýzu.

Sémantická analýza slova nám umožňuje získať všetky synonymá ku ktorémukoľvek slovu. Platí, že ak sú 2 slová, ktorých prienik synoným je nenulový, tak je vysoká pravdepodobnosť, že aj tieto 2 slová sú synonymá navzájom sebe. Vychádzame z princípu synoným, vďaka ktorej sa synonymá môžu za seba veľmi často voľne vymieňať. Preto nie je potrebné určovať synonymá ku každému záznamu kontextu aktivity, ale len k zadanému dopytu. Tieto synonymá následne porovnávame s kontextom aplikácie a opätovne hľadáme syntaktickú zhodu. Ak sa podarí nájsť, našli sme aplikáciu, ktorá je s dopytom sémanticky spojená.

Sémantická analýza nám navyše umožňuje získať aj hypernymá. Sú to slová, ktoré významovo zahŕňajú alebo obsahujú nami vyhľadávané slovo. Vďaka hypernymám môžeme nájsť spojitosť medzi kontextom aplikácie a dopytom v prípade, že vyhľadávame špecifickú časť niečoho, čím sa práve zaoberáme, napr. ak nás zaujímajú momentálne motorové vozidlá, vieme pri vyhľadávaní kamióna nájsť spojitosť s aplikáciou zameranou na motorové vozidlá, pričom motorové vozidlo je hypernymum kamióna (Obrázok 1).



Obrázok 1 Slová zdieľajúce rovnaký hypernymum, motorové vozidlo

V prípade, že stále neviem určiť s ktorou aplikáciou daný dopyt súvisí, pokúšame sa to zistiť skúmaním aktivity používateľa v počítači bez ohľadu na kontext aplikácie. To znamená, že budeme analyzovať iba fyzické aktivity ako napríklad kde používateľ klikol, čo skopíroval, bez ohľadu na text alebo akúkoľvek informáciu (okrem času), ktorá sa s danou aktivitou prenáša. Každej takejto informácií priradíme váhu a tá aplikácia, ktorá získa najvyššiu váhu je považovaná za tú, s ktorou dopyt súvisí. Toto sa však môže týkať len aplikácií, v ktorých už bola nájdená iná súvislosť, či už sémantická alebo syntaktická, inak je veľká šanca, že sa nájde prepojenie s aplikáciou s ktorou dopyt nesúvisí.

Môže sa stať, že dopyt nesúvisí s nijakou aplikáciou, ale s niečím, čo sa udialo mimo počítača. Vtedy neexistuje spojitosť medzi aplikáciou a dopytom a nami navrhovaná metóda musí skončiť tým, že neexistuje súvislosť medzi niektorou z aplikácií a dopytom. To znamená, že sa neobohatí dopyt používateľa o žiadne kľúčové slová. Inak by sa stalo, že by tie kľúčové slová neboli relevantné, keďže by vychádzali z iného kontextu ako vychádza dopyt a nami upravené vyhľadávanie by dávalo výsledky, ktoré sú v rozpore s používateľovým cieľom.

Nájdenie správneho prepojenia medzi dopytom a aplikáciou je veľmi dôležité a ak sa nám ho podarilo určiť. Získali sme širší kontext daného dopytu a môžeme ho obohatiť o nové relevantné kľúčové slová.

4.3 Obohatenie dopytu

Pre spresnenie výsledkov vyhľadávania sme sa rozhodli obohatiť dopyt o kľúčové slová kontextu aplikácie, s ktorou dopyt súvisí. Kontext aplikácie je však vyjadrený veľkým množstvom kľúčových slov a nie je možné obohatiť dopyt o všetky. Jednak nie je možné poslať vyhľadávaču na vyhľadanie také množstvo kľúčových slov a jednak nie všetky kľúčové slová sa hodia na spresnenie dopytu. Z tohto dôvodu určujeme malé množstvo najrelevantnejších kľúčových slov, ktorými obohatíme dopyt. Kľúčové slová však môžu mať inú relevanciu vzhľadom k dopytu a inú relevanciu vzhľadom ku aktuálnemu kontextu aplikácie. Preto určujeme nasledovné relevancie pre každé kľúčové slovo kontextu aplikácie:

- relevancia kľúčového slova vo vzťahu k aplikácií
- relevancia kľúčového slova vo vzťahu k dopytu

Kombinácia týchto relevancií nám určí, ktoré kľúčové slová sú najrelevantnejšie pre obohatenie používateľovho dopytu.

4.3.1 Relevancia kľúčového slova vo vzťahu k aplikácii

Každá aplikácia má svoj vlastný účel na ktorý by ju mal používateľ používať. Niektoré slúžia na pozeranie filmov, iné na programovanie a podobne. Pomocou zaznamenanej aktivity používateľa s danou aplikáciou tento účel odhaľujeme a reprezentujeme pomocou kľúčových slov. Pomocou týchto kľúčových slov vieme určiť konceptuálnu doménu aplikácie a teda aj dopyt bude patriť do rovnakej domény, ako patrí aplikácia. Ak napríklad kľúčové slová určujú, že aplikácia patrí do domény prehrávačov filmov, vieme zúžiť požadované výsledky do domény filmov.

Kľúčové slová sa snažíme získať na základe používateľovej interakcie s počítačom v priebehu celého používania aplikácie, nielen posledného časového úseku z dôvodu, že aplikácia nemení svoju doménu frekventovane. Pri určovaní relevancie kľúčových slov vychádzame z aktivity používateľa, pričom každá informácia získaná z aktivity používateľa obsahuje čas získania aktivity a typ aktivity. Neustále platí, že čím je informácia novšia, tým má väčšiu relevanciu. Vzhľadom k času sa relevancia kľúčového slova určuje podľa funkcie, ktorá umožňuje novým kľúčovým slovám pomerne rýchlo sa presadiť. Relevancia kľúčového slova však závisí aj od typu aktivity, pri ktorom bolo

zaznamenané kľúčové slovo. Každý typ má vlastnú váhu, ktorá odzrkadľuje cennosť daného typu aktivity. Celková relevancia kľúčového slova teda závisí od jeho všetkých výskytov v rámci kontextu danej aplikácií, pričom zohľadňujeme čas zaznamenania a typ aktivity, pri ktorej bolo zaznamenané.

Keďže zisťovať kľúčové slová charakterizujúce aplikáciu pri každom zadaní dopytu je z hľadiska časovej zložitosti neuskutočniteľné, musíme ich mať niekde neustále uložené, aby sme ich vedeli získať okamžite. Navrhli sme 2 metódy prehodnocovania kľúčových slov, na základe ktorých má aplikácia neustále k dispozícií presné kľúčové slová: prepočítavanie v noci offline, alebo pri každej novej aktivite používateľa automaticky prepočítať.

Prepočítavanie offline nám dá bežne dostatočne presné výsledky, pretože účel aplikácie sa mení len sporadicky, ak vôbec. Môže sa meniť pri aplikáciách ktoré slúžia na vypracovanie projektov, pričom projekty sa zvyknú vypracovávať dlhšiu dobu. V takomto prípade metóda bude síce nepresná v prvý deň začatia projektu, avšak od druhého dňa bude opäť presná, keď sa v noci znova prepočítajú relevancie kľúčových slov.

Úplné prepočítavanie relevancií kľúčových slov pri každej novej aktivite používateľa je z hľadiska náročnosti nemožné a preto je za potreby pracovať už s aktuálnymi kľúčovými slovami a ich relevanciami, ktoré už boli prepočítané. Podľa typu aktivity zisťujeme, ako nová informácia vplýva na už prepočítané kľúčové slová aplikácie a na základe tejto znalosti prepočítame novú relevanciu len tých kľúčových slov, ktoré sú obsiahnuté v práve spracovávanej aktivite. Takto vieme dostatočne zefektívniť prepočítavanie relevancie kľúčových slov pri každej novej udalosti, aby to bolo uskutočniteľné pri každej novej aktivite. Táto metóda však neberie do úvahy relevancie iných kľúčových slov a ani relevancie predošlých aktivít, v ktorých sa vyskytovali rovnaké kľúčové slová. Pracuje len s relevanciou, ktorá bola danému kľúčovému slovu už pridelená a s časom, kedy bolo ohodnotené. Na základe času odhadujeme, akú relevanciu má kľúčové slovo v tomto momente a nami vypočítanú relevanciu práve vyhodnocovanej udalosti len pripočítame k tomuto odhadu, čím získame celkovú relevanciu kľúčového slova k aplikácií v tomto momente.

4.3.2 Relevancia kľúčového slova vzhľadom k dopytu

Nie všetky kľúčové slová súvisia s dopytom, ktorý používateľ aktuálne vyhľadáva. Dopyt, ktorý používateľ vyhľadáva je vo väčšine prípadov špecifický a používateľ sa snaží nájsť mnohokrát inú informáciu, ako vyhľadával v minulosti, hoci dopyt súvisel s rovnakou aplikáciou a kontext aplikácie sa nezmenil. Z tohto dôvodu nemôžeme obohacovať dopyt len o kľúčové slová, ktoré vystihujú kontext aplikácie, ale musíme zohľadniť dopyt a obohatiť o relevantné kľúčové slová, ktoré s aktuálnym dopytom súvisia.

Kľúčové slová, ktoré súvisia s práve zadaným dopytom hľadáme len medzi aktivitou určitého posledného krátkeho obdobia, pretože predpokladáme že dopyt súvisí s niečím, čo používateľ nedávno robil v niektorej aplikácií ktorú používal. Už sme v predchádzajúcom kroku hľadali spojitosť medzi kontextom aplikácií a dopytom, preto využijeme už získané vedomosti o nájdených prepojeniach. Obohacujeme dopyt iba v prípade, že sa našla buď syntaktická, alebo sémantická spojitosť, preto musí existovať spojitosť a musíme ju v tomto momente poznať. Na základe tejto spojitosti ohodnocujeme kľúčové slová.

Pokiaľ bola nájdená syntaktická spojitosť, pokúšame sa analyzovať všetky závislosti s kľúčovým slovom, ktoré je syntakticky podobné ako niektoré kľúčové slovo z dopytu. Všetkým kľúčovým slovám, ktoré obsahovala tá aktivita, pri ktorej bolo zachytené aj syntakticky podobné kľúčové slovo, sa priradí určitá relevancia. Ak napríklad názov okna obsahoval syntakticky podobné slovo tak sa všetkým ostatným kľúčovým slovám zachyteným v názve okna pridelí rovnaká relevancia. Takisto môže byť informácia obsahujúca syntakticky zhodné kľúčové slovo len podkategóriou inej informácie, napr. ak je kľúčové slovo časťou súvislého textu v dokumente, tak je len podkategóriou prislúchajúcemu nadpisu. Keďže sa snažíme špecifikovať dopyt, predpokladáme že nadpis vyjadruje tému textu. Kľúčové slová z vyššej kategórie získajú pre to vyššiu relevanciu.

Ak bola nájdená sémantická spojitosť, znamená to, že máme nielen informáciu o tom, že je súvislosť medzi dopytom a kontextom niektorej aplikácie, ale už aj z tejto spojitosti vieme získať cenné informácie. Ak bola nájdená synonymická zhoda, vieme dopyt rozšíriť o synonymum nájdené v kontexte aplikácie, tým pádom znížime viacznačnosť dopytu. V prípade, že sa nájde hypernymum, vieme obohatiť dopyt o tento hypernymum. Tým pádom zúžime nájdené výsledky len do tej domény, ktorú určuje hypernymum. Všeobecne tak môžeme povedať, že ak sa nájde kľúčové slovo v kontexte prepojenej aplikácií, ktoré je

sémanticky prepojené s dopytom, pridelíme tomuto slovu maximálnu relevanciu, pretože vždy upresňuje zadaný dopyt.

Obohatenie dopytu nastáva pri zadaní dopytu. Počet slov, o ktoré sa má obohatiť dopyt musí byť obmedzený a zistíme ho experimentálne, pričom môžeme obohatiť dopyt buď o stále rovnaký počet najrelevantnejších dopytov, alebo sa určí spodná hranica relevancie kľúčového slova a všetky, ktoré ju presiahnu, obohatia dopyt používateľa.

Takisto je prístupná možnosť explicitnej spätnej väzby používateľa, aby si používateľ sám mohol prispôsobovať obohatenie dopytu. Poskytne nám to cennú spätnú väzbu o relevancii kľúčového slova, ktorú zohľadníme v nasledujúcom prideľovaní relevancii.

4.4 Diskusia

Nami navrhnutá metóda rozširuje dopyt na základe kontextu aplikácie, ktorá súvisí s dopytom. Jej úspech a presnosť závisí od kontextu aplikácie, ktorý sa nám podaril zistiť. Výhodou našej metódy je, že sa snaží čo najpresnejšie zisťovať kontext všetkých aplikácií, ktoré používateľ používa. Informácie, ktoré zachytávame z aktivity v týchto aplikácií sú často v prirodzenom jazyku a preto ich upravujeme na kľúčové slová, ktoré sú ľahko strojovo spracovateľné. Zisťovanie kontextu aplikácií je však ohraničené ochranou používateľovho súkromia. Našou prioritou je nezasahovať do súkromia používateľa, preto sa nebudeme snažiť získať obsah akejkoľvek aplikácie (okrem Office Word), pretože by mohol obsahovať citlivé údaje.

Dôležitým krokom je nájsť prepojenie medzi kontextom aplikácie a dopytom. Dopyt však často býva nepresne zadaný, či už bez diakritiky alebo nesprávne, pričom vyhľadávač sám opraví daný dopyt. Kvôli zle zadanému dopytu naša metóda nemusí nájsť žiadne prepojenie, preto výhodou našej metódy je, že využívame výsledky vyhľadávania na to, aby sme správne určili diakritiku a opravili dopyt.

Výhodou našej metódy je aj zohľadnenie jazyka, pretože prirodzený jazyk sa inakšie spracúva v každom jazyku. Jazyk však určujeme na krátkom úseku textu, preto sa môže často vyskytnúť nepresnosť pri určovaní jazyka. Aby sme čo najviac znížili chybovosť, rozhodli sme ohraničiť množinu jazykov na dva: slovenský a anglický.

Výhodou našej metódy je aj využitie sémantiky pri obohacovaní dopytu a hľadaním prepojenia medzi kontextom aplikácie. Vďaka nej vieme ľahšie špecifikovať viacznačnosť dopytu, pretože homonymá, ktoré majú rôzny význam, majú aj rôzne sémantické prepojenia, preto ak sa nám podarí nájsť sémantické prepojenie viacznačného kľúčového slova, vieme určiť, ktoré presne to je. Avšak stále sa nepodarilo presne zmapovať sémantické prepojenia pre všetky slová jazyka, preto sme ohraničený neúplnou sémantickou databázou, ktorá je navyše veľmi rozsiahla, preto vyhľadávanie v nej je relatívne pomalé a musí sa vhodne využívať.

5 REALIZÁCIA NAVRHNUTEJ METÓDY

Za účelom overenia navrhnutej metódy sa nám podarilo vytvoriť niekoľko softvérových nástrojov, ktoré sa dajú rozdeliť do 2 častí:

- zaznamenávanie aktivity používateľa
- analyzovanie prepojení medzi dopytom a aplikáciou

Aj keď naším prvotným cieľom bolo obohacovanie dopytu používateľa v doméne digitálnych knižníc, rozhodli sme sa našu metódu overiť všeobecnejšie na ľubovoľnej doméne. Za týmto účelom sme si zvolili analyzovanie vyhľadávania v najpoužívanejšom vyhľadávači, Google. Analyzovanie prepojení medzi dopytom a aplikáciou sme však implementovali nezávisle na tomto vyhľadávači, preto sa dá naša metóda použiť aj na doménu digitálnych knižníc jednoduchým integrovaním do prostredia Annoty[1].

Pre doménu vyhľadávača Google sme sa rozhodli z dôvodu, že je to najčastejšie používaným vyhľadávačom a používatelia v ňom vyhľadávajú najvariabilnejšie dopyty, čím si môžeme overiť metódu na analyzovanie prepojení medzi dopytom a aplikáciou na širšej množine aplikácií, ktoré málokedy súvisia s dopytom v doméne digitálnych knižníc.

5.1 Zaznamenávanie aktivity používateľa

Na získanie aktivity používateľa je potrebné realizovať modul, ktorý bude zaznamenávať používateľovu aktivitu na počítači. Podľa navrhnutej metódy na zachytenie aktivity používateľa sa nám podarilo zrealizovať 3 samostatné logovacie systémy:

- Tabber – zaznamenáva aktivity na rozhraní operačného systému
- Annota-extension – zaznamenáva aktivitu v prehľadávači
- Wordik – zaznamenáva aktivitu v Office Word

[1] Annota, http://annota.fiit.stuba.sk/

Tieto 3 moduly sme si vybrali za účelom overenia nami navrhnutých metód. Je možné ich rozšíriť o ďalšie logovacie moduly, ktoré budú odosielať zaznamenanú aktivitu vo vhodnom tvare na náš server.

Za pomoci týchto 3 moduloch vieme odchytávať mnohé aktivity, avšak len v prirodzenom, resp. nespracovanom jazyku. Z tohto dôvodu sú potrebné na spracovanie aktivít ešte nasledujúce moduly:

- Ruby-stemmer
- LemmatizerWebService
- CompactLanguageDetection

Vďaka týmto modulom vieme spracovať zachytenú používateľovu aktivitu a získať kľúčové slová, s ktorými sa jednoducho manipuluje. Spracovanie jazyka prebieha automaticky na každej zaznamenanej aktivite, preto ak by sa pridal ďalší logovací modul, informácie o zachytenej aktivite v prirodzenom jazyku by sa automaticky spracovali na kľúčové slová.

5.1.1 Tabber

Tabber je jednoduchá .NET aplikácia, ktorá sníma používateľovu aktivitu v každej aplikácii na vrstve operačného systému. Túto schopnosť jej dáva možnosť importovať windowsové knižnice. Vďaka ním sa dokáže priamo napojiť na udalosti, ktoré vyvolávajú aplikácie keď chcú komunikovať s operačným systémom. Odchytávame nasledovné udalosti:

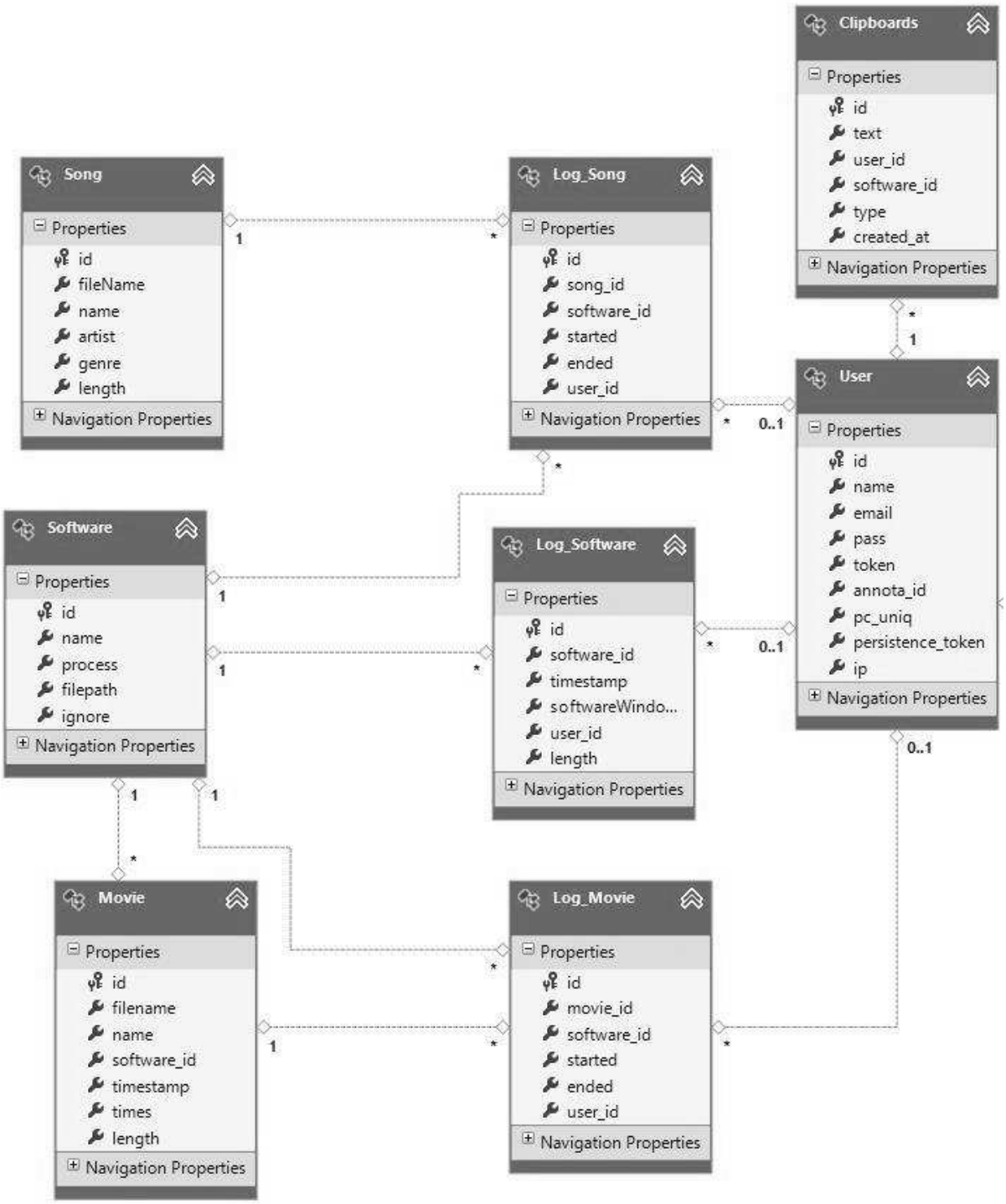
- prepnutie aplikácie
- prehrávanie filmu / piesne
- skopírovanie / prilepenie textu

Prepnutie aplikácie zachytávame z dôvodu, že je pre nás dôležité zistiť, v ktorej aplikácií sa používateľ nachádzal a aký dlhý čas. Navyše vieme zistiť postupnosť prepínania aplikácií a teda aj v ktorej aplikácii sa používateľ nachádzal pred tým, ako zadal dopyt. Táto udalosť nám okrem času poskytuje aj iné informácie: názov aplikácie a názov okna aplikácie, do ktorej sa používateľ prepol.

Prehrávanie filmov a piesní vieme zachytiť vďaka systému, akým multimediálne prehrávača fungujú. Súbor, ktorý momentálne prehrávajú si uzamknú na vrstve operačného systému, aby ho nemohol nikto meniť počas prehrávania. Vieme zistiť všetky súbory, ktoré sú uzamknuté iným aplikáciám pre upravovanie. Podľa prípony následne určíme o aký typ súboru ide a ak je to multimediálny súbor, pokúšame sa zistiť ešte presnejšie informácie, pretože názov súboru nemusí mnoho hovoriť. Pri pesničkách vieme zistiť vďaka ID3 značkovaniu aj názov interpreta, názov piesne a žáner, do ktorého patrí. Pri filmoch sa snažíme zistiť len presný názov filmu. Tieto informácie zisťujeme kvôli tomu, že používateľ používa počítač často za účelom prehrávania multimédií, pričom vyhľadáva rôzne dopyty súvisiace s práve prehrávaným multimédiom, napr. titulky, hodnotenia a podobne.

Skopírovanie textu zachytávame vďaka sledovaniu aktivity schránky. Keď používateľ skopíruje časť textu, tento text zaznamenáme a rovnako aj aplikáciu, v ktorej daný text skopíroval, podobne ako pri prilepení textu do aplikácie. Získavame takto veľmi cenné informácie o texte používateľa, ktorý považuje za hodnotný, v ktorejkoľvek aplikácií, pretože tento úsek textu sa chystá použiť ešte v inej aplikácií.

Obrázok 2 znázorňuje dátový model, ktorý slúži na ukladanie kontextu a podrobne zobrazuje jednotlivé údaje, ktoré sa ukladajú pre jednotlivé typy aktivity.



Obrázok 2 Dátový model slúžiaci na zachytenie aktivity používateľa v prostredí Windows

5.1.2 Annota-extension

Annota-extension je rozšírenie do prehľadávača Mozilla Firefox, ktoré umožňuje zaznamenávanie používateľovej aktivity v priestore prehľadávača. Toto rozšírenie už bolo implementované tímom Annota za účelom sledovania používateľa v doméne internetových

knižníc. Keďže sme rozšírili overenie na všetky aplikácie, nielen internetové knižnice, rozšírili sme toto rozšírenie o zaznamenávanie aktivity v každej webovej aplikácií. V prehliadači máme stránky, ktoré slúžia na prezentáciu obsahu a aplikácie, ktorých účelom je vykonávanie určitej aktivity, napr. objednanie tovaru. V našej realizácií považujeme za webovú aplikáciu aj prezentovanie webového obsahu, pričom za určujúce pre jej identifikáciu považujeme názov domény. Zaznamenávame podobné udalosti ako pri desktopových aplikáciách:

- zmena aktívnej stránky
- kliknutie na odkaz
- skopírovanie a označenie textu
- prehrávanie multimediálneho obsahu v doméne Youtube

Zmena aktívnej záložky a kliknutie na odkaz nám slúži na získanie pohybu používateľa v internetovom prehliadači. Pri zmene aktívnej stránky navyše získavame informáciu o identifikačnom znaku záložky, takže vieme určiť, či používateľ otvoril novú stránku na novej záložke alebo aktuálnu stránku nahradil novou stránkou.

5.1.3 Wordik

Wordik je rozšírenie do Microsoft Office Word, ktoré zaznamenáva používateľovu aktivitu v tejto aplikácií. Predpokladáme, že centrom používateľovho záujmu je text, ktorý sa práve upravuje, preto sa zamierame práve naň a keď používateľ opustí aplikáciu, zaznamenávame:

- aktuálny odsek
- odsek nad a pod aktuálnym odsekom
- nadpis aktuálneho odseku a aj vyššie úrovne aktuálneho nadpisu

Nadpisy zvyknú vystihovať tému nasledujúceho textu, preto sú pre nás veľmi cenné. Aktuálny odsek a jeho okolie nám poskytujú dostatok informácií o aktuálne písanom texte, avšak sú tvorené v prirodzenom jazyku a tak z textu nevieme získať zopár kľúčových slov, ktoré vyjadrujú daný text najpresnejšie.

5.1.4 Moduly pre spracovanie prirodzeného jazyka

Ruby-stemmer a LemmatizerWebService slúžia na stemovanie slov, pričom Ruby-stemmer slúži pre stemovanie anglických slov a LemmatizerWebService pre stemovanie slovenských slov.

CLD slúži na získavanie jazyka, pričom podporuje aj slovenčinu. Vracia kód jazyka a dôveryhodnosť tohto odhadu. Keďže určujeme jazyk na základe len krátkeho textu, tento modul nie vždy uhádne jazyk správne. Pre zjednodušenie určujeme slovenčinu v prípadoch, že vráti český, slovenský alebo poľský jazyk a angličtinu v prípadoch, že vráti iný jazyk.

5.2 Analýza prepojení medzi dopytom a aplikáciou

Hľadanie súvislosti medzi kontextom aplikácie a dopytom používateľa sa nehľadá u používateľa, ale na serveri. Hľadanie sa začína pri zadaní dopytu a prebieha na pozadí, aby nespomaľovalo používateľa.

Na pozadí používame server Annota, ktorý bol už implementovaný za účelom umožnenia používateľovi pridávať k ľubovoľným dokumentom poznámky a vytvárať záložky (Bieliková et al., 2013). Slúži taktiež na vyhľadávanie vedeckých článkov, čo je dôvodom toho, že sme naše overenie implementovali práve do tohto systému. Aj keď nakoniec neobohacujeme dopyt pri vyhľadávaní v systéme Annota, je možné integrovať našu metódu aj do tohto vyhľadávania, pretože nami navrhovaná metóda funguje cez aplikačné rozhranie. Prijíma JSON požiadavky s dopytom a vracia aplikácie, ktoré súvisia s dopytom.

Sémantická analýza bola realizovaná za pomoci modulu Wordnet. Wordnet je lexikálna databáza anglického jazyka. Podstatné mená, slovesá a prídavné mená sú spojené do synsetov, pričom každý jeden vyjadruje určitý koncept. Synsety sú prepojené konceptuálno-sémantickými a lexikálnymi vzťahmi. Wordnet je voľne dostupný a používame ho na zistenie synoným a hyperrnýn dopytu. Wordnet však poskytuje lexikálnu databázu len pre anglický jazyk, preto v tejto implementácii zatiaľ nie sme schopní hľadať sémantické prepojenia pre slovenské slová.

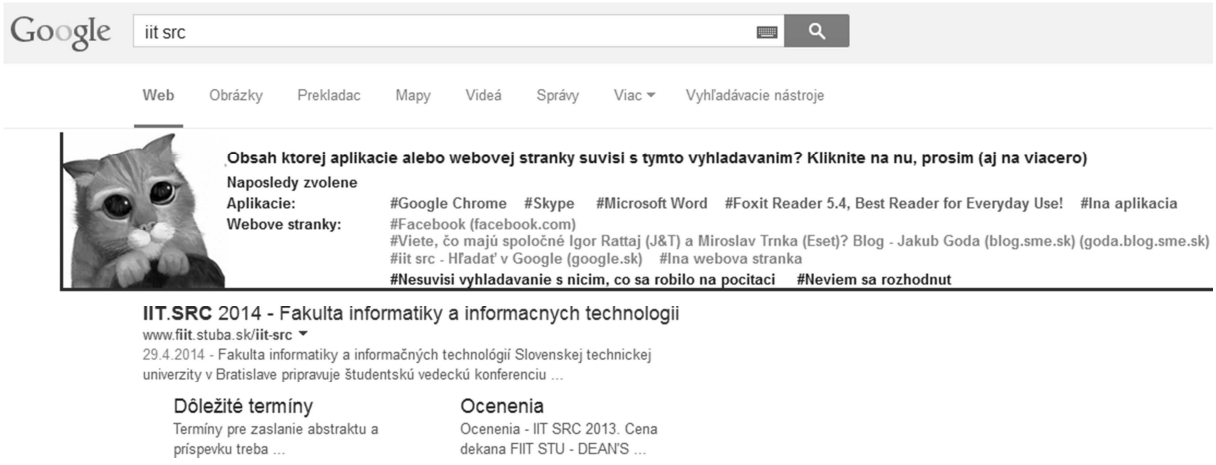
Algoritmus, ktorým sa hľadá prepojenie medzi aplikáciou a dopytom sme implementovali na základe nami navrhnutej metódy. Jednotlivé kroky sú nasledovné:

1. Získanie kontextov všetkých aplikácií používaných posledný časový úsek
2. Hľadanie syntaktickej zhody medzi dopytom a kontextom aplikácií
3. Získanie synoným a hyperným každého kľúčového slova dopytu
4. Hľadanie syntaktickej zhody medzi kontextom aplikácií a synonymami a hypernymami dopytu
5. Ohodnotenie aktivity používateľa s ohľadom len na typ aktivity
6. Vypočítanie pravdepodobnosti, že aplikácia súvisí s dopytom, na základe syntaktickej zhody, sémantickej blízkosti a aktivity používateľa vzhľadom len na typ aktivity

6 EXPERIMENTÁLNE VYHODNOTENIE

Overenie navrhnutej metódy sme vykonali formou otvoreného experimentu v prostredí vyhľadávača Google. Experiment sme vykonávali na vzorke 5 ľudí, ktorých sme požiadali, aby označovali podľa vlastného uváženia ktoré z aplikácií súvisia s aktuálnym dopytom. Experimentátori boli študenti informatiky a architektúry.

Za týmto účelom sme pozmenili vyhľadávaciu stránku Google, na ktorej sme umožnili spojiť dopyt s aplikáciou jednoduchým kliknutím na jej názov. Obrázok 3 znázorňuje používateľské rozhranie integrované do vyhľadávača Google. Pre lepšiu prehľadnosť sme používateľovi zobrazili iba 4 najdlhšie používané desktopové aplikácie a 4 najdlhšie používané webové aplikácie, v ktorých sme predpokladali, že používateľ nájde tú aplikáciu, ktorá súvisí s dopytom. Pre reprezentovanie všetkých možností mohol používateľ označiť, že aplikácia nesúvisí s ničím čo sa vykonávalo na počítači, resp. že daná aplikácia nie je v zozname.



Obrázok 3 Používateľské rozhranie pre spájanie aplikácií s dopytom

Experiment prebiehal počas 6 týždňov. Zameriavali sme na sledovanie používateľa v prirodzenom prostredí pri plnení bežných úloh. Experiment bol preto počas celej doby nekontrolovaný a prebiehal u každého používateľa na jeho vlastnom počítači. Počas celej doby prebiehajúceho experimentu mali nainštalovaný a aktívny nami vytvorený logovací modul, pričom súhlasili s tým, že ich aktivita bude zaznamenávaná a spracovaná na účely tohto experimentu. Používatelia explicitne prepojili 470 dopytov s aplikáciami.

V našej práci sme sa zamerali na overenie 2 oblastí:

- existuje prepojenie medzi dopytom a aplikáciou
- vieme určiť, ktorá aplikácia súvisí s dopytom

Zamerali sme sa na tieto 2 oblasti, pretože aby sme obohacovali dopyt o kontext niektorej aplikácie, musíme vedieť presne určiť, ktorá to je aplikácia.

6.1 Existuje prepojenie medzi dopytom a aplikáciou

Aby sme mohli zistiť, ktorá aplikácia súvisí s dopytom, musíme najprv zistiť, či vôbec existuje prepojenie medzi dopytom a aplikáciou, s ktorou dopyt súvisí. K overeniu hypotézy, že existuje prepojenie medzi dopytom a aplikáciou, ktorú používateľ nedávno používal využívame explicitnú spätnú väzbu, ktorú nám poskytli používatelia zapojení do experimentu.

Používatelia mali za úlohu označiť aplikáciu, ktorej obsah súvisí s aktuálnym vyhľadávaním. Tým, že nejakú aplikáciu označili, explicitne potvrdili, že existuje prepojenie medzi dopytom a niektorou z aplikácií, ktorú nedávno používali. **Chyba! Nenalezen zdroj odkazů.**Tabuľka 2 zobrazuje počet dopytov, ktoré prepojili s niektorou aplikáciou a počet dopytov, ktoré nesúvisia s aktivitou na počítači.

	Webové aplikácie	**Desktopové aplikácie**	**Neexistuje prepojenie**	**Spolu**
Počet	178	218	56	472
Percentuálny podiel	37,71%	46,16%	11,86%	100%

Tabuľka 2 Explicitné prepojenia dopytu s aplikáciou

Experiment ukázal, že skoro 90% dopytov súvisí s niektorou aplikáciou, ktorú používateľ nedávno používal. Znamená to, že používatelia vyhľadávajú predovšetkým dopyty prepojené s aktivitou na počítači a teda v 88,14% dopytov existuje súvislosť s niektorou aplikáciou.

Našou snahou je odstrániť nutnosť explicitne prepájať dopyt s aplikáciami a získavať prepojenie automaticky z aktivity používateľa. Za týmto účelom sme navrhli

a implementovali 2 metódy, ktoré sa chystáme overiť práve na používateľom prepojených dopytoch. Vybrali sme si túto množinu z dôvodu, že používateľ označil, že obsah určitej aplikácie je prepojený s dopytom, preto by sme mali v týchto prípadoch nájsť prepojenie aj našimi metódami.

Prvou metódou je syntaktická analýza, ktorá hľadá v kontexte aplikácie prepojenej s dopytom syntaktickú podobnosťou s dopytom. Brali sme do úvahy kontext získaný v stanovenom časovom intervale. My sme vo všetkých experimentoch použili časový interval 15 minút. Je to čas, ktorý sme určili na základe pozorovaní tak, aby vo väčšine prípadov zahŕňal ucelený kontext. Sú však situácie, keď by tento čas mal byť dlhší alebo kratší a bolo by vhodné aj s týmto parametrom pracovať dynamicky. Vzhľadom na rozsah tohto projektu sme pri experimentovaní uvažovali konštantný interval. Tabuľka 3**Chyba! Nenalezen zdroj odkazů.** zobrazuje výsledky tohto experimentu.

	Slovenské dopyty			**Anglické dopyty**			**Spolu**		
	Nájdené	**Všetky**	**%**	**Nájdené**	**Všetky**	**%**	**Nájdené**	**Všetky**	**%**
Desktopové aplikácie	3	14	21,4	91	175	52,0	94	189	49,7
Webové aplikácie	38	52	73,1	81	122	66,4	119	174	68,4
Všetky aplikácie	41	66	62,2	172	297	57,9	213	363	58,7

Tabuľka 3 Výsledky syntaktickej analýzy hľadania prepojenia medzi dopytom a prepojenou aplikáciou

Ukázalo sa, že za pomoci len syntaktickej analýze dokážeme nájsť 58,7% prepojení, pričom pri webových aplikáciách je úspešnosť o 20% vyššia. Dôvodom je, že webové aplikácie poskytujú lepšie informácie o aktuálnom dokumente, pretože názov webových dokumentov väčšinou vystihuje aj obsah dokumentu a často sa mení, pričom názov okna pri desktopových aplikáciách väčšinou znázorňuje len názov súboru a názov aplikácie a nemení sa často.

Takisto sme zistili, že sa v anglickom jazyku vyhľadáva vo väčšej miere, pri webových aplikáciách 70% dopytov bolo v anglickom jazyku a pri desktopových aplikáciách až 95%. Môže to byť dôsledkom toho, že podpora pre desktopové aplikácie je väčšinou kvalitnejšia v anglickom jazyku a keďže všetci experimentátori dosahujú v angličtine aspoň úroveň B2, je im pohodlnejšie vyhľadávať v tomto jazyku. Dopyty v slovenskom jazyku slúžili hlavne na získanie lokálnych informácií týkajúcich sa miest a inštitúcií na Slovensku, preto súviseli väčšinou s aktivitou vo webových aplikáciách. Dá sa z toho usúdiť, že ak je dopyt v Slovenčine a používateľ plynule ovláda tento jazyk, je veľmi vysoká pravdepodobnosť, že súvisí s webovou aplikáciou a nie s desktopovou aplikáciou.

Druhá metóda je sémantická analýza, ktorá hľadá sémanticky blízke slová dopytu v kontexte prepojenej aplikácie. Táto metóda sa však dala aplikovať len na vzorku anglických dopytov vzhľadom k tomu, že sémantickú analýzu podporujeme len pre tento jazyk. Tabuľka 4 zobrazuje výsledky hľadanie sémantickej podobnosti medzi dopytom a kontextom aplikácie.

	Nájdené	**Všetky**	**%**
Desktopové aplikácie	61	175	34,9
Webové aplikácie	19	122	15,6
Všetky aplikácie	80	297	26,9

Tabuľka 4 Výsledky sémantickej analýzy hľadania prepojenia medzi dopytom a prepojenou aplikáciou

Dokázali sme nájsť sémantické prepojenie medzi dopytom a aplikáciou v 26,9% prípadoch.

Keďže sémantická analýza porovnáva vždy iné kľúčové slová s kontextom aplikácií ako syntaktická analýza, hľadali sme súvislosť medzi dopytom a kontextom aplikácií skombinovaním sémantickej a syntaktickej analýzy. Tabuľka 5 zobrazuje výsledky hľadania prepojenia skombinovaním týchto 2 metód.

	Slovenské			Anglické			Spolu		
	Úspešné	**Všetky**	**%**	**Úspešné**	**Všetky**	**%**	**Úspešné**	**Všetky**	**%**
Desktopové aplikácie	3	14	21,4	106	175	60,6	109	189	57,7
Webové aplikácie	38	52	73,1	81	122	66,4	119	174	68,4
Všetky aplikácie	41	66	62,1	187	297	63,0	228	363	62,8

Tabuľka 5 Výsledky hľadania prepojenia medzi dopytom a prepojenou aplikáciou skombinovaním metód

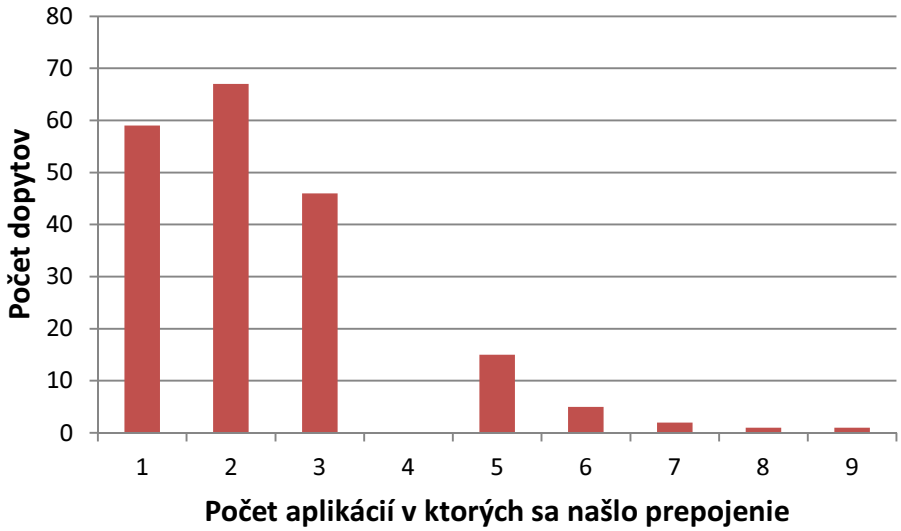
V predchádzajúcom experimente sme našli len 15,6% sémantických prepojení. Z toho dôvodu nám skombinovanie obidvoch metód nepomohlo zvýšiť úspešnosť hľadania prepojení. Aj keď sa nám podarilo nájsť v 34,2% prípadoch sémantické prepojenie medzi desktopovými aplikáciami, výrazne nám to nezvýšili úspešnosť (8%). Pri oboch typoch aplikácie sme teda zistili, že medzi syntakticky podobnými aplikáciami a sémanticky blízkymi aplikáciami je veľký prienik 81,3%.

Ukázali sme, že v 62,8% prípadoch vieme nájsť prepojenie medzi dopytom a aplikáciou, ktorý s dopytom súvisí. Najčastejšia je táto súvislosť syntaktická a teda používatelia hľadajú rovnaké slová, ako vidia v danej aplikácií. Nepodarilo sa nám odhaliť viac prepojení z dôvodu, že sme sa kvôli ochrane súkromia účastníkov experimentov snažili nezaznamenávať citlivé údaje ani nijaký obsah aplikácie, okrem aplikácie Microsoft Office.

6.2 Určenie aplikácie súvisiacej s dopytom

Určiť, ktorá aplikácia súvisí s dopytom je možné len v prípade, že sme našli sémantickú blízkosť alebo syntaktickú podobnosť, inak by hrozilo s veľkou pravdepodobnosťou, že vyberieme nesprávnu aplikáciu. Prepojenie sme dokázali nájsť správne v 62,8% prípadoch. Bližšia analýza však ukázala, že sa bežne nájde prepojenie aj s viacerými aplikáciami, ako

len s jednou. Obrázok 4 znázorňuje počet dopytov vzhľadom k počtu aplikácií, pri ktorých sa v danom dopyte našlo prepojenie, či už syntaktické alebo sémantické.



Obrázok 4 Počet dopytov vzhľadom k počtu aplikácií v ktorých sa našlo prepojenie k danému dopytu

Vo väčšine prípadov sa nájde prepojenie v 1-3 aplikáciách naraz. Keďže používateľ vždy označil, že dopyt nesúvisí so všetkými aplikáciami, je potrebné určiť, ktoré sú tie správne. Zisťujeme to na základe už nájdených syntaktických a sémantických prepojení a na používateľovej aktivite v jednotlivých aplikáciách. K dispozícii máme rôzne za posledný časový úsek indikátory pre každú aplikáciou, ktorá syntakticky, alebo sémanticky súvisí s aktuálnym dopytom:

- aktívny čas v aplikácii (podľa poradia)
- počet prepnutí do danej aplikácie (podľa poradia)
- počet skopírovaní v danej aplikácii (podľa poradia)
- počet aktívnych aplikácií medzi dopytom a poslednou aktivitou v aplikácií
- počet sémantických ostemovaných prepojení
- počet sémantických prepojení v prirodzenom jazyku
- počet syntaktických ostemovaných prepojení
- počet syntaktických prepojení v prirodzenom jazyku

Podľa poradia znamená, že sa aplikácie zoradili od najlepšej po najhoršiu podľa určitého indikátora a najlepšie umiestnená dostala 5 bodov, pričom každá ďalšia o bod menej.

Na základe týchto indikátorov určujeme, ktorá aplikácia súvisí s dopytom. Nie každý indikátor má však rovnakú relevanciu, preto je potrebné určiť váhu každého indikátora, akou vplýva na určenie aplikácie, ktorá je prepojená s dopytom. Váhu sme zisťovali po ukončení experimentu na získaných dátach, pričom sme sa snažili priradiť váhu jednotlivým indikátorom tak, aby aplikácie, ktoré používateľ označil, že súvisia s dopytom mali najvyššiu relevanciu a tie, ktoré s dopytom nesúvisia, mali najnižšiu relevanciu.

Váhy sme určovali za pomoci strojového učenia, konkrétne evolučného algoritmu. Fitnes sme počítali ako počet prípadov, keď používateľom určená aplikácia dosiahla najvyššiu relevanciu a odpočítali sme jednotlivé poradia umiestnenia používateľom zvolenej aplikácie, keďže čím bola aplikácia relevantnejšia, tým bola vyššie umiestnená. Tabuľka 6 zobrazuje váhy jednotlivých indikátorov získaných za pomoci evolučného algoritmu.

Indikátor	**Váha**
Aktívny čas v aplikácii (podľa poradia)	6.71
Počet prepnutí do danej aplikácie (podľa poradia)	1.58
Počet skopírovaní v danej aplikácii (podľa poradia)	1.46
Počet aktívnych aplikácií medzi dopytom a poslednou aktivitou v aplikácií	3.70
Počet sémantických ostemovaných prepojení	3.89
Počet sémantických prepojení v prirodzenom jazyku	0.53
Počet syntaktických ostemovaných prepojení	1.03
Počet syntaktických prepojení v prirodzenom jazyku	-1.10

Tabuľka 6 Váhy jednotlivých indikátorov

Experiment ukázal, že niektoré indikátory majú omnoho vyššiu relevanciu ako iné. Najvýraznejší vplyv dosiahla aktívna dĺžka aplikácie, ktorá má veľký náskok pred ostatnými indikátormi. Dá sa teda usúdiť, že aplikácia, ktorá súvisí s dopytom sa dá určiť

hlavne na základe toho, koľko času v nej používateľ strávil za poslednú dobu. Ak teda používateľ intenzívne pracuje v niektorej aplikácií, je veľká pravdepodobnosť, že bude aj jeho vyhľadávanie súvisieť s touto aplikáciou. Výrazný vplyv má aj počet nájdených sémantických prepojení. Je to hlavne dôsledkom toho, že len v malom množstve aplikácií sa našla sémantická podobnosť, preto ak sa našla, je vysoká pravdepodobnosť, že je to práve s prepojenou aplikáciou. Rovnako vysokú relevanciu má počet aktívnych aplikácií medzi dopytom a posledným prepnutím sa z aplikácie. Znamená to, že dopyt výrazne súvisí s aplikáciou, z ktorej sa používateľ naposledy prepol.

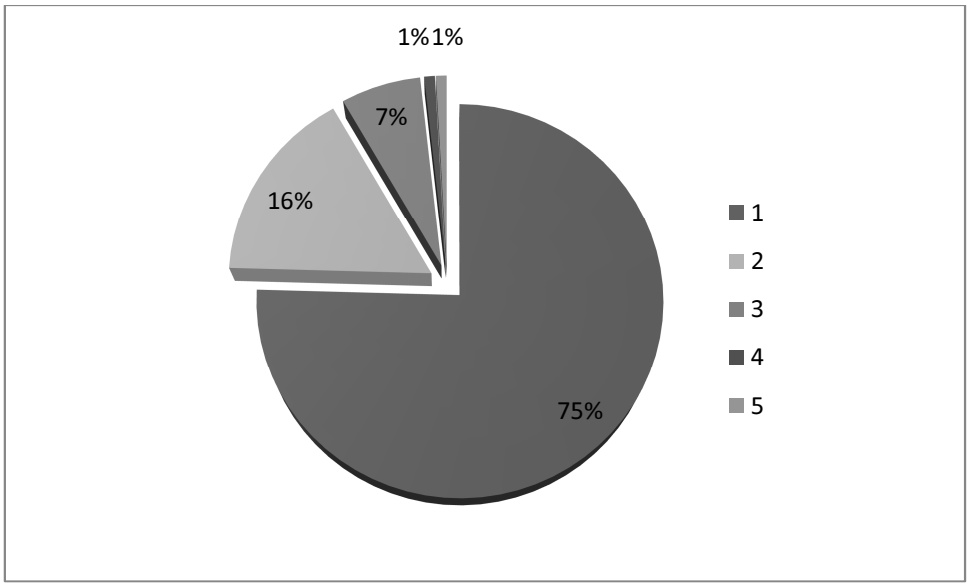
Zaujímavým výsledkom je negatívna váha pre počet syntaktických prepojení v prirodzenom jazyku. Môže to byť následkom toho, že syntaktická podobnosť je najčastejšie nájdeným prepojením medzi aplikáciou a dopytom a nachádza sa skoro vo všetkých aplikáciách, teda aj u tých, ktoré s dopytom nesúvisia.

Našou úlohou je nájsť aplikáciu, ktorá súvisí s dopytom. Na základe ohodnotenia aktivity používateľa v danej aplikácií podľa vyššie spomínaných indikátorov sme skúmali, koľko aplikácií, ktoré používateľ označil, že súvisia s aktuálnym dopytom, dosiahnu najvyššiu relevanciu. Tabuľka 7 zobrazuje výsledky nášho experimentu, ktorý mal za úlohu nájsť prepojenie medzi aplikáciou a dopytom.

Správne určené aplikácie	**Počet nájdených prepojení**	**%**
172	228	75,4

Tabuľka 7 Výsledky určovania aplikácie, ktorá súvisí s dopytom

Dokázali sme správne určiť, ktorá aplikácia súvisí s dopytom na základe indikátorov v 75,4% prípadoch. Ak sa teda nájde prepojenie medzi aplikáciou a dopytom, vieme pomerne presne zistiť na základe stanovených indikátorov, ktorá aplikácia súvisí s dopytom. Obrázok 5 počet umiestnení prepojenej aplikácií na danej priečke.



Obrázok 5 Počet dopytov, v ktorých dosiahla prepojená aplikácia dané umiestnenie

Obrázok 5 ukazuje, že v 16% prípadoch dosiahla lepšie umiestnenie len jedna iná aplikácia ako tá, ktorú označil používateľ. Môže to znamenať chybu našej metódy, avšak tak isto to môže označovať aplikácie, ktoré s dopytom súvisia, avšak používateľ ich neoznačil. Pri bližšom skúmaní aplikácií, ktoré sa umiestnili na prvom mieste sme zistili, že práve toto druhé vysvetlenie vyzerá pravdepodobnejšie. Používatelia totiž ani raz neoznačili, že dopyt súvisí s viac ako 1 aplikáciou. Znamená to, že dopyt súvisí v niektorých prípadoch aj s viac, ako jednou aplikáciou. Ak sa nám podarí odhaliť viacero aplikácií, prepojených s dopytom, získame ešte širší kontext vyhľadávania, ktorý využijeme pri obohacovaní dopytu.

7 ZÁVER

V tejto práci sme navrhli a implementovali metódu na získanie kontextu aplikácie sledovaním používateľovej aktivity v počítači. V metóde sa zaoberáme pozorovaním používateľa v celom prostredí počítača, zameriavame sa pritom na aplikáciu Microsoft Office Word a webový prehliadač Mozilla Firefox, zohľadňujeme však aj všetky desktopové aplikácie. Prirodzený jazyk, v ktorom získavame informácie spracúvame na kľúčové slová, ktoré vystihujú kontext aplikácií. Vďaka týmto kľúčovým slovám vieme určiť kontext aplikácie za ľubovoľný časový úsek.

Navrhli a implementovali sme metódu na hľadanie prepojenia medzi kontextom aplikácie a dopytom. Medzi kľúčovými slovami dopytu a kľúčovými slovami kontextu aplikácie hľadá súvislosť, ktorou môže byť buď syntaktická podobnosť, alebo sémantická blízkosť. Znamená to, že hľadáme prekryv rovnakých slov, hoci aj v inom tvare a prekryv významovo príbuzných slov, napr. synoným. Táto metóda nám však často našlo prepojenie medzi viacerými aplikáciami naraz, preto sme navrhli a implementovali metódu, ktorá z nich vyberie tú, ktorá je prepojená s aktuálnym dopytom. Za týmto účelom sme navrhli 8 indikátorov relevancie aplikácie k dopytu, ktoré určia najrelevantnejšie aplikácie vzhľadom k aktuálnemu kontextu vyhľadávania a teda tie, ktoré súvisia s aktuálnym dopytom.

Nami navrhnuté riešenie sme overili neriadeným experimentom v prostredí vyhľadávača Google. Zapojili sme do neho 5 študentov STU, ktorých úlohou bolo zaznamenávať aktivitu na svojom počítači pomocou nami vytvorených modulov a v prípade, že vyhľadávajú niečo na internete, označiť aplikáciu, ktorá súvisí s aktuálnym dopytom. Používatelia takto priradili 396 dopytov k aplikácií, ktorá s nimi súvisela. Zistili sme na základe týchto prepojení, že skoro 90% dopytov súvisí s aktivitou, ktorú používateľ vykonával v počítači.

Našou snahou však je určovať prepojenie medzi aplikáciou a dopytom automaticky, bez nutnosti používateľovej spätnej väzby. Nami navrhnuté metódy, ktoré slúžia tomuto cieľu sme overili práve na používateľom prepojených dopytoch, kde sme vedeli, ktorá aplikácia je prepojená s aktuálnym dopytom. Dokázali sme nájsť prepojenie v 58,7% prípadoch

pomocou syntaktickej analýzy a v prípade anglických dopytov v 26,9% prípadoch pomocou sémantickej analýzy. Skombinovaním týchto metód sme dosiahli 62,8% úspešnosť nájdenia prepojenia medzi dopytom a aplikáciou. Určovať, ktorá aplikácia súvisí s aktuálny dopytom je možné len ak bola nájdená súvislosť s aplikáciou, preto sme metódu na určovanie, ktorá aplikácia súvisí s dopytom, overovali na množine aplikácií, v ktorých sa už našla súvislosť. Rôzne indikátory, ktoré určovali relevantnosť aplikácie vzhľadom k dopytu, mali rôzne relevancie. Relevancie sme určili pomocou evolučného algoritmu. Pomocou týchto indikátorov sa nám podarilo správne určiť 75% aplikácií a v 15% prípadoch používateľom určená aplikácia skončila na druhom mieste, pričom podrobnejšie analýza ukázala, že v mnohých prípadoch súvisel dopyt v týchto prípadoch s viacerými aplikáciami, nielen s tou, ktorú používateľ označil.

Podarilo sa nám celkovo ukázať, že používateľ vyhľadá hlavne dopyty, ktoré súvisia s prácou na počítači a že vo veľkej miere vieme správne určiť aplikáciu, ktorá súvisí s dopytom len na základe zaznamenávania používateľovej aktivity v prostredí počítača.

V ďalšej práci sa môžeme zaoberať skúmaním používateľovej aktivity vo vybranej špecifickej doméne, s využitím znalostí o doméne. Mohli by sme sa zamerať na doménu digitálnych knižníc, kde predpokladáme časté prepojenie medzi písaním textov a vyhľadávaním zdrojov. Keďže je možné zachytávať širokú škálu aktivít pri písaní textov, napr. úroveň nadpisov, podčiarknutý text, zaoberali by sme využitím tejto rôznorodosti pre určovanie kontextu aktuálne písaného textu a obohacovanie dopytu.

Inou doménou, na ktorú by sme sa mohli zamerať, je doména hudby. Už teraz vieme získať podrobne pesničky, ktoré používateľ počúva. Mohli by sme sa venovať odporúčaniu hudby, pričom by sme zohľadňovali jednak jeho celkový vkus a jednak piesne, ktoré momentálne počúva. Zamerať by sme sa mohli pri tom na negatívnu implicitnú spätnú väzbu, ktorá by nám vedela rýchlo usmerniť hudobný vkus používateľa.

LITERATÚRA

1. Attar, R. and Fraenkel, A.S., 1977. Local Feedback in Full-Text Retrieval Systems. *Journal of the ACM*, 24(3), pp.397–417.

2. Aula, A., Khan, R.M. and Guan, Z., 2010. How does search behavior change as search becomes more difficult? In *Proceedings of the 28th international conference on Human factors in computing systems - CHI '10*. New York, USA: ACM Press, p. 35.

3. Bateman, S., Teevan, J. and White, R.W., 2012. The search dashboard. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12*. New York, New York, USA: ACM Press, p. 1785.

4. Bieliková, M., Ševcech, J., Holub, M. and Móro, R., 2013. Annota – annotating the documents in the domain of digital libraries. In *DATAKON '13: Proc. of the Annual Database Conf. VŠB-TU*. Ostrava, pp. 143–152.

5. Dey, A.K., Abowd, G.D., Brown, P.J., Davies, N., Smith, M. and Steggles, P., 1999. Towards a Better Understanding of Context and Context-Awareness. In *Proceedings of the 1st international symposium on Handheld and Ubiquitous Computing*. London, UK: Springer-Verlag, pp. 304 – 307.

6. Dou, Z., Song, R. and Wen, J.-R., 2007. A large-scale evaluation and analysis of personalized search strategies. In *Proceedings of the 16th international conference on World Wide Web - WWW '07*. New York, USA: ACM Press, p. 581.

7. Eirinaki, M. and Vazirgiannis, M., 2003. Web mining for web personalization. *ACM Transactions on Internet Technology*, 3(1), pp.1–27.

8. Fitzpatrick, L. and Dent, M., 1997. Automatic feedback using past queries. *ACM SIGIR Forum*, 31(SI), pp.306–313.

9. Jansen, B.J., Spink, A. and Saracevic, T., 2000. Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing & Management*, 36(2), pp.207–227.

10. Jiang, D., Leung, K.W.-T. and Ng, W., 2011. Context-aware search personalization with concept preference. In *Proceedings of the 20th ACM international conference on Information and knowledge management - CIKM '11*. New York, New York, USA: ACM Press, p. 563.

11. Joachims, T., Granka, L., Pan, B., Hembrooke, H., Radlinski, F. and Gay, G., 2007. Evaluating the accuracy of implicit feedback from clicks and query reformulations in Web search. *ACM Transactions on Information Systems*, 25(2), p.7–es.

12. Jones, R. and Klinkner, K.L., 2008. Beyond the session timeout. In *Proceeding of the 17th ACM conference on Information and knowledge mining - CIKM '08*. New York, New York, USA: ACM Press, p. 699.

13. Kim, J.Y., Cramer, M., Teevan, J. and Lagun, D., 2013. Understanding how people interact with web search results that change in real-time using implicit feedback. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management - CIKM '13*. New York, New York, USA: ACM Press, pp. 2321–2326.

14. Kramár, T., 2014. Utilizing Lightweight Semantics for Search Context Acquisition in Personalized Search. *Information Sciences and Technologies Bulletin of the ACM Slovakia*, p.6.

15. Lagun, D., Sud, A., White, R.W., Bailey, P. and Buscher, G., 2013. Explicit feedback in local search tasks. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval - SIGIR '13*. New York, New York, USA: ACM Press, p. 1065.

16. Leung, K. and Lee, D., 2010. Deriving concept-based user profiles from search engine logs. *Knowledge and Data Engineering, IEEE Transactions*, 22(7), pp.969–982.

17. Peska, L. and Vojtas, P., 2013. Negative implicit feedback in e-commerce recommender systems. In *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics - WIMS '13*. New York, New York, USA: ACM Press, p. 1.

18. Pirolli, P. and Card, S., 1995. Information foraging in information access environments. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '95*. New York, New York, USA: ACM Press, pp. 51–58.

19. Polzonetti, A., 2010. User-centric mobile services: context provisioning and user profiling. In *Proceedings of the 11th Annual International Digital Government Research Conference on Public Administration Online: Challenges and Opportunities (dg.o '10)*. Digital Government Society of North America, pp. 122–130.

20. Qiu, F. and Cho, J., 2006. Automatic identification of user interest for personalized search. In *Proceedings of the 15th international conference on World Wide Web - WWW '06*. New York, New York, USA: ACM Press, p. 727.

21. Ryan, N., Pasco, J. and Morse, D., 1998. Enhanced Reality Fieldwork: the Context-Aware Archaeological Assistant. In V. Gaffney, M. Leusen, & S.- Exxon, eds. *Computer Applications in Archaeology*. Oxford: Tempus Reparatum, pp. 182–196.

22. Schilit, B., Adams, N. and Want, R., 1994. Context-Aware Computing Applications. In *1994 First Workshop on Mobile Computing Systems and Applications*. IEEE, pp. 85–90.

23. Speretta, M. and Gauch, S., Personalized Search Based on User Search Histories. In *The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)*. IEEE, pp. 622–628.

24. Sun, J., Zeng, H.-J., Liu, H., Lu, Y. and Chen, Z., 2005. CubeSVD. In *Proceedings of the 14th international conference on World Wide Web - WWW '05*. New York, New York, USA: ACM Press, p. 382.

25. Teevan, J., 2008. How people recall, recognize, and reuse search results. *ACM Transactions on Information Systems*, 26(4), pp.1–27.

26. White, R.W. and Morris, D., 2007. Investigating the querying and browsing behavior of advanced search engine users. *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '07*, p.255.