Záměr týmového projektu

Studijní program	Informatika – Softwarové a datové inženýrství (PV, VŠ) Informatika – Softwarové systémy (FM)
Typ projektu	Výzkumný projekt (NPRG070)
Studenti-řešitelé	Patrik Veselý, Bc., Vít Škrhák, Bc., František Mejzlík, Bc.
Supervizor	doc. RNDr. Jakub Lokoč, Ph.D. (KSI)
Konzultant	Mgr. Ladislav Peška, Ph.D. (KSI)
Název a téma projektu	Rozšíření state-of-the-art vyhledávacího systému SOMHunter

Motivace projektu

Tento projekt reaguje na potřeby úspěšného vyhledávacího nástroje *SOMHunter* [1] vyvíjeného výzkumným týmem SIRET na Katedře softwarového inženýrství. Reaguje jak na nutnost jeho pravidelného rozšiřování, tak i na nutnost podrobně analyzovat dosavadní fungování systému.

V současnosti je prioritou integrace nových textových dotazovacích přístupů a implementace nového uživatelského rozhraní. Tyto rozšíření jsou součástí příprav nástroje pro letošní soutěže *Video Browser Showdown (VBS)* [2] a *Lifelog Search Challenge (LSC)* [3], kde minulý rok nástroj SOMHunter obsadil první a druhé místo.

Modely zpětnovazebního učení, textové modely a SOM mapy, jsou tři pilíře celého systému. Proto existuje potřeba vylepšení nejen dosavadních textových modelů, ale i algoritmů pro zpětnovazební učení. Samotné porovnání těchto algoritmů je náročný úkol, protože za tímto účelem nemůže existovat benchmark dataset a potřebujeme k tomu lidské uživatele, což je nepraktické. Lidské operátory se snažíme nahradit automatickým modelem uživatele, abychom mohli aproximovat výkon algoritmů a strategií hledání bez nutnosti rozsáhlých uživatelských studií. Za tímto účelem bude vytvořen nástroj pro sběr dat o poskytování zpětné vazby a spuštění rozsáhlé uživatelské studie, která poskytne data o tom, jak uživatelé vybírají snímky z displeje. Zároveň poskytne data pro vytvoření automatického modelu uživatele.

Dále bude vytvořen framework k automatickému ověření nových algoritmů pro vyhledávání pomocí zpětné vazby za použití nově získaného automatického uživatele. Ten bude následně využit k hledaní optimálního nastavení systému SOMHunter a k vývoji dalších budoucích

rozšíření zpětnovazebních algoritmů. Všechny poznatky o výběru reálných a simulovaných uživatelů budou shrnuty do publikace v žurnálu *Multimedia Tools and Applications (MTAP)* [4], jež měl v roce 2019 impact factor 2.313.

Druhým vědeckým výstupem bude zpracování podrobné analýzy výkonu SOMHunteru (i za pomoci nasbíraných logů) z poslední soutěže LSC 2020, kde nástroj SOMHunter obsadil druhé místo. Díky tomuto úspěchu jsme byli pořadatelem LSC přizváni k vytvoření publikace pro speciální vydáni žurnálu MTAP.

Popis a cíle projektu

Celý projekt se skládá z několika menších podcílů, které jako celek tvoří významný posun celého vyhledávacího systému SOMHunter a to nejen ve smyslu softwarového díla, ale také posun na základě vědeckých poznatků, které vzejdou z provedených analýz a experimentů. Projekt také počítá se vznikem dvou publikací, které budou zaslány do recenzovaného žurnálu s IF.

A. **Implementace textového dotazování pro libovolný podregion snímku** videa, což umožní formulovat dotazy schopné zachytit i pozičnı́ vztahy entit ve videu. Současná verze systému umožňuje popis pouze snímku jako celku.

B. **Relokace dotazů**, která dá uživateli možnost iterativně vylepšovat jednotlivé komponenty jeho kombinovaného dotazu na základě okamžité zpětné vazby pro dané modality dotazu. Hypotéza je, že přesnější jednotlivé složky povedou k větší celkové přesnosti dotazu.

C. **Reimplementace současného webového uživatelského rozhraní** za pomocí lépe vyhovujícího frameworku. Tento bod souvisí s předešlými kroky, protože nové uživatelské rozhraní bude počítat i s rozšířeným dotazováním. Mimo to, výsledkem bude výrazně čitelnější a přehlednější kód front-endu, který umožní snadnější nástup do projektu pro budoucí členy týmu.

D. **Implementace softwaru pro sběr “relevance feedback” dat**, který bude následně využit k rozsáhlé uživatelské studii. Tento software bude vytvořen z aktuální verze nástroje. Bude prezentovat zjednodušený postup vyhledávání a implementovat vhodný mechanismus tvorby logů.

E. **Analýza logů a připravení simulace reálného uživatele** na základě nasbíraných dat, pomocí softwaru z předchozího bodu. Tyto simulace slouží k automatickému nastavení nástroje SOMHunter. Výstupem bude analýza výběru obrázků z displeje a algoritmus simulace tohoto výběru pomocí heuristik a strojového učení. Poznatky z této studie budou shrnuty do **publikace** a zaslány do impaktovaného žurnálu MTAP.

F. **Analýza logů ze soutěže LSC 2020** společně s porovnáním, jak by si systém vedl, kdyby již tehdy používal nový model CLIP. Výstupem budou data, grafy a další informativní vizualizace výkonu systému, na základě kterých budeme rozhodovat o budoucím vývoji systému SOMHunter. Na základě provedené analýzy bude sepsána **publikace** shrnující výsledky provedené analýzy a porovnání. Tato publikace bude zaslána také do žurnálu.

G. **Integrace text-to-video modelu *CLIP*** [5], který dle našich předběžných měření vykazuje vyšší přesnost na našem test datasetu než momentálně využívaný model W2VV++ [6]. Tato zásadní změna by měla přinést výrazně vyšší výkonnost systému, zároveň ale vyžaduje zásahy do celkové architektury systému.

H. Posledním bodem projektu je **zobecnění textového dotazování** na dotazování buď pomocí textu nebo pomocí obrázku dodaného z libovolného zdroje (typicky bitmapa ze schránky systému). Přičemž uživatel může typ jednotlivě umístěných dotazů (z bodu A) libovolně kombinovat. To dovolí uživatelům popsat i koncepty, na které se těžko hledají slova nebo koncepty, které implementovaný text-to-image model neumí dobře umístit.

Technické řešení

Stávající systém SOMHunter je rozdělen do tří klíčových částí.

- **Jádro** vyvíjené v jazyce *C++*, které implementuje veškerou logiku a použité modely. Také udržuje stav uživatelských hledání. Jádro je samostatná jednotka schopna použití v libovolné aplikaci, která jádro bude používat přes standardní C++ třídu.
- **Mezivrstva vystavující API** jádra ve formě HTTP(S) serveru za pomoci *Node.js* [7].
- **Webové uživatelské rozhraní**, které komunikuje se serverovou mezivrstvou (tedy tranzitivně i s jádrem) a umožňuje koncovým uživatelům snadno používat systém.

Vzhledem k rozmanitosti cílů se během práce na tomto projektu řešitelé setkají s několika technologiemi.

A. Rozšíření dotazování bude vyžadovat zobecnění rozhraní jádra, kde již dotaz nebude jeden řetězec, ale komplexní struktura. Pro samotné zpracování těchto dotazů bude zapotřebí mít přístup k vyextrahovaným rysům pro různé podregiony, které se následně budou využívat na základě umístění uživatelských dotazů. Tato část bude prováděna zejména pomocí standardního C++. Také bude třeba provést příslušné rozšíření serverové mezivrstvy (Node.js) a uživatelského rozhraní, aby mohl uživatel pohodlně tyto dotazy zadávat (viz C).

B. Integrace relokovaných dotazů bude vyžadovat novou komponentu v uživatelském rozhraní, které bude koncový uživatel schopen používat. Také bude nutné rozšíření jádra o podporu rychlého dotazování na menší SOM mapy a menší rychle dostupné shrnující displeje pro větší počet dotazů. Implementace tohoto bodu bude pravděpodobně probíhat souběžně s nově vznikajícím uživatelským rozhraním (viz C).

C. Implementaci nového uživatelského rozhraní bude provedena pomocí frameworku *Ember.js* [8], protože pro tento typ silně interaktivního rozhraní nabízí vhodnou vnitřní architekturu, která vede ke kratšímu a čitelnějšímu kódu. Během implementace bude také vytvořeno rovnou nové rozhraní pro zadávání dotazů na podregiony (viz A).

D. Implementace softwaru pro sběr dat o výběrech z displeje bude provedena úpravou stávající verze systému SOMHunter. Do jádra systému bude doplněna podpora pro více paralelních uživatelských připojení, zobrazení hledaného cíle, návrh a implementace speciálního podrobného logování uživatelských akcí a ladící nástroje pro ověření validity dat.

E. K analýze dat nasbíraných nástrojem z bodu D a vytvoření modelu simulovaného uživatele budou využity jazyky R nebo Python. Framework pro simulaci hledání pomocí simulovaného uživatele bude implementován nad jádrem SOMHunter.

F. Analýza a evaluace námi pořízených logů během soutěže LSC 2020 bude prováděna především s pomocí jazyka Python a jeho běžných knihoven (např. numpy, matplotlib). Protože v současnosti vzniká nový standardizovaný formát těchto logů (společný pro všechny týmy soutěžící na VBS a LSC) může být výstupem i nástroj, jež bude příslušné vizualizace provádět pro libovolné logy nástrojů dodržujících tento formát. Výsledky a zjištění analýzy budou shrnuty do vědeckého článku, vytvořeného dle předepsaného stylu vydavatelem žurnálu MTAP.

G. Integrace CLIP modelu bude asi technicky nejobtížnějším úkolem. Bude zapotřebí výrazně upravit fungování jádra a vyřešit několik zásadních otázek typu – jakým způsobem se bude textový dotaz umisťovat do vektorového prostoru? Oproti současnému řešení, kde se pro daný dotaz sečetly předpočítané vektory a přenásobily maticí, bude zapotřebí textový dotaz zobrazit několika hlubokými modely, které musí někde běžet.

Bude se muset do jádra v C++ integrovat prostředí pro běh těchto modelů, což pravděpodobně bude implikovat nedostatek hlavní paměti a vyústí to v netriviální úpravu práce s daty, nad kterými systém počítá. Případně sáhneme po datech částečně mapovaných na disk, nebo některé části systému odsuneme na sdílený server, kde potřebné části vystavíme jako službu.

Také je možné, že hluboké rysy CLIP modelu nebudou vhodné pro zpětnovazební modely a proto bude nutno pracovat částečně i nad rysy z modelu W2VV++.

H. Zobecnění dotazování bude vyžadovat úpravu rozhraní společně s jádrem systému. Bude nutno zavést ještě obecnější reprezentaci dotazu a na dotazy podané ve formě obrázku bude zapotřebí pro tyto obrázky vyextrahovat jeho rysy a následně napojit na stávající textový model.

Spolupráce ve výzkumném týmu

Celý projekt bude řešen v rámci výzkumné skupiny SIRET na Katedře softwarového inženýrství, se kterou řešitelé již delší dobu pravidelně spolupracují. Supervizor tohoto projektu, **doc. Jakub Lokoč**, bude spolupracovat na řešení odborných problémů, bude dohlížet na průběh projektu a průběžně kontrolovat kvalitu výstupů. Řešitelé se budou se supervizorem (a dle potřeby i s ostatními členy výzkumného týmu SIRET) scházet na pravidelných konzultacích, které budou probíhat každý týden (případně dle potřeby a dohody).

Bližší spolupráce s ostatními členy týmu lze shrnout takto:

A. Na implementaci rozšíření dotazování bude pracovat **František Mejzlík**.

B. Na implementaci relokace dotazu bude pracovat **Patrik Veselý**.

C. Implementaci uživatelského rozhraní provede **František Mejzlík** a **Vít Škrhák**. Uživatelské testování nového rozhraní (společně s již hotovými rozšířeními) budou provádět **všichni řešitelé**.

D. **Patrik Veselý** naimplementuje nástroje pro sběr dat a ladící nástroje pro data vytvoří **Vít Škrhák**.

E. Analýzu dat výběrů uživatelů provede **Vít Škrhák** s odbornou pomocí **doc. Jakuba Lokoče, Dr. Marty Vomlelové** a **Dr. Ladislava Pešky**, kteří se budou podílet i na publikaci. Na implementaci frameworku se budou podílet **Vít Škrhák** a **Patrik Veselý**.

F. Analýzu logů z LSC provedou řešitelé **František Mejzlík** a **Patrik Veselý** pod metodickým vedením **doc. Jakuba Lokoče.** Na tvorbě publikace budou spolupracovat s konzultantem **Dr. Ladislavem Peškou** a **doc. Jakubem Lokočem**.

G. Na integraci CLIP modelu budou pracovat **všichni řešitelé** společně s pomocí **Mgr. Tomáše Součka**.

H. Úpravu architektury dotazování provedou **Vít Škrhák** a **Patrik Veselý**.

Předběžný průběh prací

Předběžný plán prací pro **9 měsíců** ([1 2 3 4 5 6 7 8 9]) je předběžně stanoveno takto:

Bod A (*Implementace textového dotazování pro libovolný podregion snímku*)
- Řešitel **František Mejzlík**
- Měsíce **[1 - - - - - - - -]**

Bod B (*Relokace dotazů*)
- Řešitel **Patrik Veselý**
- Měsíce **[1 - - - - - - - -]**

Bod C (*Reimplementace současného webového uživatelského rozhraní*)
- Řešitelé **František Mejzlík** a **Vít Škrhák**
- Měsíce **[1 2 3 - - - - - -]**

Bod D (*Implementace softwaru pro sběr "relevance feedback" dat*)
- Řešitelé **Vít Škrhák** a **Patrik Veselý**
- Měsíce **[- 2 3 4 - - - - -]**

Bod E (*Analýza logů a připravení simulace reálného uživatele, publikace*)
- Řešitelé **Vít Škrhák** a **Patrik Veselý**
- Měsíce **[- - - 4 5 6 - - -]**

Bod F (*Analýza logů ze soutěže LSC 2020, publikace*)
- Řešitelé **František Mejzlík** a **Patrik Veselý**
- Měsíce **[- - 3 4 5 - - - -]**

Bod G (*Integrace text-to-video modelu CLIP*)
- Řešitelé **Patrik Veselý**, **Vít Škrhák** a **František Mejzlík**
- Měsíce **[- - - - 5 6 7 8 -]**

Bod H (*Zobecnění textového dotazování*)
- Řešitelé **Vít Škrhák** a **Patrik Veselý**
- Měsíce **[- - - - - - - 8 -]**

Bod Z (*Finalizace, testování a dokončení dokumentací*)
- Řešitelé **Patrik Veselý**, **Vít Škrhák** a **František Mejzlík**
- Měsíce **[- - - - - - - - 9]**

Reference

[1] Kratochvíl, M., Veselý, P., Mejzlík, F. and Lokoč, J., 2020, January. **SOM-hunter: Video browsing with relevance-to-som feedback loop.** In *International Conference on Multimedia Modeling* (pp. 790-795). Springer, Cham.

[2] Rossetto, L., Gasser, R., Lokoc, J., Bailer, W., Schoeffmann, K., Muenzer, B., Soucek, T., Nguyen, P.A., Bolettieri, P., Leibetseder, A. and Vrochidis, S., 2020. **Interactive video retrieval in the age of deep learning-detailed evaluation of VBS 2019.** *IEEE Transactions on Multimedia*.

[3] Gurrin, C., Schoeffmann, K., Joho, H., Leibetseder, A., Zhou, L., Duane, A., Dang-Nguyen, D.T., Riegler, M., Piras, L., Tran, M.T. and Lokoč, J., 2019. **Comparing Approaches to Interactive Lifelog Search at the Lifelog Search Challenge (LSC2018).** *ITE Transactions on Media Technology and Applications*, *7*(2), pp.46-59

[4] **Multimedia Tools and Applications** https://www.springer.com/journal/11042

[5] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. and Krueger, G., 2021. **Learning transferable visual models from natural language supervision.** *arXiv preprint arXiv:2103.00020*.

[6] Li, X., Xu, C., Yang, G., Chen, Z. and Dong, J., 2019, October. **W2VV++ fully deep learning for ad-hoc video search.** In *Proceedings of the 27th ACM International Conference on Multimedia* (pp. 1786-1794).

[7] **Node.js** https://nodejs.org/en/

[8] **Ember.js** https://emberjs.com/