Agentúra na podporu výskumu a vývoja
Mýtna 23, P.O.BOX 839 04, 839 04 Bratislava 32

Formulár ZK

Záverečná karta projektu

Názov projektu **Antiplagiatorská analýza netextových dát**

Evidenčné číslo projektu **APVV–0469–12**

Zodpovedný riešiteľ **Mgr. Ján Grman, PhD.**

Príjemca **Slovenská technická univerzita v Bratislave**

Názov pracoviska, na ktorom bol projekt riešený

1. Ústav elektrotechniky, Fakulta elektrotechniky a informatiky, STU Bratislava
2.
3.
4.
5.

Názov a štát zahraničného pracoviska, ktoré spolupracovalo pri riešení

1.
2.
3.

Udelené patenty/podané patentové prihlášky, vynálezy alebo úžitkové vzory, ktoré sú výsledkami projektu

1.
2.
3.

Najvýznamnejšie publikácie (knihy, články, prednášky, správy a pod.) zhrňujúce výsledky projektu – uveďte aj publikácie prijaté do tlače

1.
2.
3.
4.
5.

Uplatnenie výsledkov projektu

Výsledky projektu sú uplatniteľné v dvoch typoch projektov. Prvou možnosťou je tvorba

nástroja pre vyhľadávanie obrázkov. Teda na základe vstupného obrázku sa prehľadáva databáza obrázkov a očakávaným výsledkom je zoznam identických a/alebo podobných obrázkov. Miera podobnosti definuje relevanciu. Druhou a mierne jednoduhšou možnosťou je implementácia submodulu pre systém pre odhaľovanie plagiátov. V tomto prípade je prioritou hľadanie nadprahovej podobnosti. Práve toto použitie je vzhľadom na skúsenosti s vyhľadávaním plagiatorstva v texte najviac reálne.

CHARAKTERISTIKA VÝSLEDKOV

Súhrn výsledkov riešenia projektu a naplnenia cieľov projektu v slovenskom jazyku (max. 20 riadkov)

Projekt je zameraný na prácu s netextovou informáciou. Ako prvú sme vyriešili úlohu dolovania a predspracovania netextových dát. Vytvorili sme referenčný korpus PDF prác (z Centrálneho registra záverečných prác a z internetu). Implementovali sme algoritmy pre extrakciu obrázkov a ich ukladanie. Navrhli sme a získali špeciálnu databázu pre účely ukladania dát (obrazobvé informácie) a metadát (deskriptívne popisy) pre účely experimentov a ich vyhodnocovania. Navrhli sme a experimentálne overili systém hromadného spracovania veľkých objemov dát. Analyzovali sme množstvo existujúcich algoritmov pre úlohy normalizácie a unifikácie dát. Navrhli sme modifikácie algoritmov, nové spôsoby ich spájania a vykonali overovacie experimenty a hodnotenie. V závere projektu sa nám podarilo overiť takmer celú linku spracovania dokumentov až po reprezentácie použiteľné pre antiplagiátorskú analýzu (v existujúcom antiplagiátorskom systéme). Niektoré algoritmy a postupy by pre účely samostatného modulu bolo potrebné reimplementovať mimo prostredia MATLAB. Experimenty boli realizované na širšej skupine obrazov aká sa predpokladá v reálnom nasadení. Boli zisťované hranice použiteľnosti algoritmov. Sporné výsledky sú dosahované pri porovnávaní malých obrazov (pod 100 bodov). Obmedzením dolnej hranice spoľahlivosť stúpa a metódy sú stále vyhovujúce reálnemu použitiu. Redukcia objemu dát ktorá tým vzniká, umožní realizovať porovnávanie v časoch, ktoré je možné použiť v reálnom prostredí a jeho kritériach pre odozvu systému. Za originálny prínos je možné považovať návrh a implementáciu binárnej LAT metódy a na nej postavenej metóde kvantifikácie miery zhody. Na umelo vytvorených obrazoch sa táto metóda ukazuje ako jediná vhodná.

Súhrn výsledkov riešenia projektu a naplnenia cieľov projektu v anglickom jazyku (max. 20 riadkov)

The project focuses on working with non-text information. First, we have solved the task of extraction and pre-processing of non-textual data. We have created a reference database of PDF files (from the Central Repository of Theses and form the Internet). We have implemented algorithms for extracting images and storing them. We designed a special database for data (images) and metadata (descriptors) storage and for experimental evaluation. We have designed and experimentally verified a processing link for data in large scale. We analyzed a number of existing algorithms for task of standardization and data unification. We proposed algorithm modifications and new ways to use them in one link and then performed verification experiments and evaluation. At the end of the project, we were able to verify processing data link from input document up to the representations usable for anti-plagiarism analysis (in the existing anti-plagiarism system). Some algorithms and procedures would need to be re-implemented for the purposes of a separate module outside the MATLAB environment. The experiments were carried out on a wider range of images as assumed in real deployment. Algorithms of usability limits have been investigated. Unusable results were obtained when comparing very small pictures (below 100 points). By limiting the lower size, reliability increases and the methods are still fit for real usage. Reducing the volume of data will allow us to make comparisons at times that can be used in the real environment. The original benefit can be considered as the design and implementation of the binary LAT method and the method of quantifying the degree of conformity based on this method modification. On artificial images, this method appears to be the only one appropriate.

Svojím podpisom potvrdzujem, že údaje uvedené v záverečnej karte sú pravdivé a úplné a súhlasím s ich zverejnením.

Zodpovedný riešiteľ	**Štatutárny zástupca príjemcu**
Mgr. Ján Grman, PhD.	prof. Ing. Robert Redhammer, PhD.
V Bratislave 23.10.2017	V Bratislave 23.10.2017
.....	.....
podpis zodpovedného riešiteľa	podpis štatutárneho zástupcu príjemcu