BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2023
http://excel.fit.vutbr.cz

Extrakce informací z webových dokumentů pomocí grafových neuronových sítí

Josef Katrňák*

Abstrakt

Cílem práce je využití metod strojového učení pro klasifikaci specifických částí obsahu webových stránek. Jako vstup pro experimenty je použita datová sada veřejně dostupných webových stránek obsahujících stránky on-line prodávaných produktů.

Webové stránky jsou nejprve reprezentovány pomocí vizuální reprezentace. Tato reprezentace je poté použita pro trénování modelů, jejichž architektura je založena na grafových neuronových sítích.

V práci je experimentováno s různými architekturami sítí a parametry jejich trénování. Nejlepší natrénované modely dosahují přesnosti 97,55% s F1 skóre 0,9737. Pro účely extrakce je dosažena úspěšnost nalezení požadované informace až 92,58%.

Výsledky práce jsou dosaženy bez použití textových vlastností a jsou založeny čistě na vizuálních informacích z webové stránky. Výhodou je extrakce informací nezávislá na struktuře a jazyku webové stránky.

*xkatrn00@stud.fit.vutbr.cz, *Fakulta informačních technologií Vysokého učení technického v Brně*

1. Úvod

I když existují automatické metody získávání určitých a pro uživatele užitečných částí webových stránek, stále se v oblasti extrakce a klasifikace těchto částí objevují problémy. Tyto problémy pramení v neustále rostoucím množství stránek, které se kategoricky neliší, avšak samotná struktura a vlastní obsah stránek může být značně odlišný. Možností jak s těmito problémy bojovat je použití strojového učení pro vytvoření flexibilního nástroje umožňujícího rychlé získání klíčových informací z jakékoliv webové stránky dané kategorie.

Tato práce se zaměřuje na automatickou extrakci strukturovaných dat z webových stránek. Pro experimenty je v práci zvolena úloha nalezení informací o produktu z e-shopových webových míst. Cílem je extrahovat tyto informace nezávisle na konkrétním webovém místě a jeho struktuře. Problém je převeden na klasifikaci jednotlivých částí webové stránky a evaluace je provedena pomocí metrik jako je přesnost, skóre F1 a přesnost nominace.

Existující nástroje pro získání strukturované informace z nestrukturovaného vstupu se nazývají *web wrappery*. Jejich nevýhodou je potřeba adaptovat se na konkrétní šablonu stránky a tedy úzké navázání na specifickou sadu webových stránek. Dalším problémem je potřeba pravidelných úprav a aktualizací wrapperů při změně webových stránek.

Přístup této práce je založen na získání strukturované reprezentace webového dokumentu pomocí nástroje FitLayout a následné využití této reprezentace pro trénování modelu hluboké neuronové sítě. Tento model má dvě části: embedder, který slouží k zachycení vlastností webové stránky a klasifikátor pro určení zda jednotlivé části obsahují požadovanou informaci. Embedder se skládá z vícevrstvé grafové neuronové sítě, zatímco klasifikátor představuje klasickou lineární neuronovou vrstvu.

Při experimentech nejlepší natrénované modely dosahovaly přesnosti 97,55% a F1 skóre 0,9737. Nominační přesnost, tedy úspěšnost nalezení požadované informace byla až 92,58%. Grafové neuronové sítě se ukazují jako vhodné pro zachycení vlastností jednotlivých částí webových stránek a vztahů mezi nimi. Na výsledný embedding je možné využít lineární klasifikátor a díky tomu extrahovat požadované informace z webových stránek. Výhodou tohoto přístupu je vysoká nezávislost na konkrétní struktuře webové stránky a tedy i nezávislost na konkrétním webovém místě.

2. Extrakce informací z webových stránek

Pro úkol extrakce klíčových informací z webových stránek byla zvolena datová sada The Klarna Product Page Dataset [1]. Tato datová sada obsahuje stránky on-line prodávaných produktů na rozdílných webových místech. Datová sada obsahuje offline snímek 51,701 produktových stránek od 8,175 různých prodejců z 8 různých zemí (Německo, USA, Velká Británie, Finsko, Rakousko, Švédsko, Norsko a Nizozemsko) v MHTML formátu. Pro každou stránku je přímo v HTML kódu anotováno 5 požadovaných informací: název, cena, hlavní obrázek produktu a tlačítka pro přidání a přesměrování do košíku. Na **Obrázku 1** je vidět příklad takové stránky s vyznačenými požadovanými informacemi.

3. Reprezentace webové stránky

Každá webová stránka je poté reprezentována pomocí nástroje FitLayout [2]. Stránky jsou nejprve vyrenderovány, poté je provedena segmentace pro seskupení vizuálních oblastí stránky. Na **Obrázku 2** je vidět výstup segmentace, kde jednotlivé vizuální oblasti jsou organizovány ve stromové struktuře. Každá oblast má sadu vlastností jako je například úroveň zanoření, její pozice na stránce, barevné vlastnosti nebo vlastnosti textu. Takto reprezentované oblasti jsou uloženy ve formátu JSON a následně převedeny na grafy, kde každá oblast představuje uzel grafu a vazby mezi uzly znázorňují vztah rodiče a potomka ze stromové struktury. Listy stromu jsou reprezentovány speciálními uzly, které jsou cílem klasifikace. Tyto uzly mohou obsahovat požadovanou informaci.

4. Model neuronové sítě

Model vytvořený pro tuto práci se skládá ze dvou částí, jak je vidět na **Obrázku 3**. První část se nazývá embedder a vstupem do této části je uzel grafu reprezentující jednu oblast webové stránky a graf samotný. Embedder se skládá z grafových neuronových vrstev. Počet těchto vrstev je jedním z parametrů modelu a je možné s ním experimentovat. Stejně tak je parametrem velikost jednotlivých vrstev. Na výběr je ze tří architektur embedderu: grafová konvoluční síť (GCN) [3], grafová síť založená na *attention* mechanismu (GAT) [4] nebo síť typu vícevrstvý perceptron (MLP). Výstupem embedderu je embedding o zvolené velikosti, který reprezentuje vlastnosti daného uzlu vzhledem ke kontextu grafu a požadované úloze.

Pro klasifikaci uzlů v grafových neuronových sítí se využívá kromě vlastností jednotlivých uzlů i lokální a globální informace o struktuře grafu. Grafové neuronové sítě iterativně propagují informace z jednotlivých uzlů do jejich sousedních uzlů, což umožňuje uzlům získávat informace o svém okolí. Při použití klasické sítě typu vícevrstvý perceptron se kontext uzlu nebere v potaz a pracuje se pouze s vlastnostmi jednotlivých uzlů. Ve srovnání v **Tabulce 1** je vidět, že grafové neuronové sítě si vedli lépe než MLP.

Druhá část modelu, která navazuje na embedder, je klasifikátor. Ten má za úkol na základě poskytnutého embeddingu klasifikovat daný uzel. Výstupem je tedy vektor o velikosti počtu cílových informací, jehož interpretací je možné zjistit jestli daný uzel obsahuje požadovanou informaci a případně o jakou informaci se jedná. Klasifikátor je implementován jako klasická lineární neuronová vrstva.

5. Experimenty a závěr

Experimenty se zaměřují na dvě metriky. Jedna z nich je celková přesnost klasifikace uzlů a odpovídající skóre F1. Tato metrika, ale může být zkreslená kvůli velkému množství uzlů, které žádnou informaci neobsahují. Je tedy možné, že klasifikátor predikuje ve velkém množství majoritní třídu, ale i tak dosahuje poměrně velké úspěšnosti. Jelikož se v zadané úloze na každé stránce vyskytuje požadovaná informace maximálně jednou byla zvolena alternativní metrika nominační přesnosti. Ta říká v kolika procentech případů dokáže model najít cílovou informaci, pokud se na stránce opravdu nachází.

Při experimentech se ukázalo, že nevyváženost této úlohy (většina klasifikovaných uzlů neobsahuje informaci) může být problém. Ve snaze bojovat s tímto problémem bylo vyzkoušeno více ztrátových funkcí, které využívají principu váhování pro vzorky obsahující minoritní třídy. V **Tabulce 2** je vidět srovnání jednotlivých ztrátových funkcí.

Dále bylo experimentováno s různými nastaveními parametrů modelu a trénování sítě. Sada konfigurovatelných parametrů obsahuje 21 položek včetně použité sítě, počtu vrstev a jejich velikostí, nastavení ztrátové funkce nebo optimalizátoru trénování.

V **Tabulce 2** jsou také shrnuty nejlepší zatímní dosažené výsledky. S GAT architekturou a ztrátovou funkcí Cross Entropy Loss byla dosažena přesnost **0,9755** a skóre F1 **0,9737**. Stejná architektura, ale s váhovanou Cross Entropy Loss docílila přesnosti nominace **0,9285**.

Předmětem pokračující práce je experimentování s nastavením a kombinací dalších parametrů a možnosti přidání textových vlastností k uzlu.

Literatura

[1] Alexandra Hotti, Riccardo Sven Risuleo, Stefan Magureanu, Aref Moradi, and Jens Lagergren. The klarna product page dataset: A realistic benchmark for web representation learning. *CoRR*, abs/2111.02168, 2021.

[2] Martin Milicka and Radek Burget. Information extraction from web sources based on multi-aspect content analysis. In Fabien Gandon, Elena Cabrio, Milan Stankovic, and Antoine Zimmermann, editors, *Semantic Web Evaluation Challenges*, pages 81–92, Cham, 2015. Springer International Publishing.

[3] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *CoRR*, abs/1609.02907, 2016.

[4] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks, 2018.