An Approach for Collecting Real Estate Development News

Matheus de Oliveira Salim, Vinicius Ferreira Salgado, Wladmir Brandão*
Pontifical Catholic University of Minas Gerais
Belo Horizonte, Minas Gerais, Brazil
wladmir@pucminas.br

Daniel Henrique Mourão Falci, Fernando Silva Parreiras*
LIAISE
Belo Horizonte, Minas Gerais, Brazil
{daniel.falci, fparreiras}@liaise.com.br

I. THE APPROACH

Fig. 1. The architecture of the proposed approach

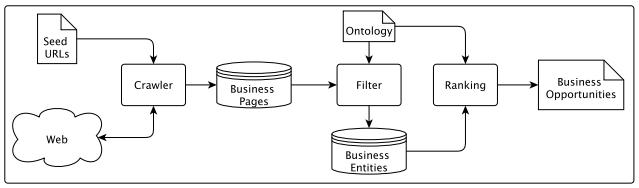

Figure 1 presents the architecture of our approach to collect, classify and rank real estate development news from the Web. We highlight that the crawler component of our approach firstly uses a set of seed URLs to collect real estate development news, storing them in a business pages corpus, re-feeding itself from URLs contained in the collected pages. The crawling flow used by our approach is similar to a previous work reported in [1].

Second, the filter component extracts textual features, i.e., page title, page description, and page text, from the real estate development news corpus. For this, it filters valid pages by removing pages from invalid URLs, duplicated, empty, redirected and private pages. Then, it performs stop words removal, stemming, and removal of punctuation, accents and special characters in order to increase the quality of the textual features. Finally, it exploits a real estate development ontology to extract entities from the textual features.

Third, the ranking component uses supervised algorithms to learn a ranking model in order to provide an ordered list with the most relevant business opportunities extracted from the Web.

II. Preliminary Experiments and Results

To evaluate our approach, we run experiments to answer the following research questions: i) how effective is our approach to collect and filter real estate development news? ii) which textual features provide better filtering performance? The evaluation of the entity recognition and ranking strategies will be carried out in future work.

Particularly, we use a dataset composed of 419 real estate development news, previously collected from the Web and labeled by experts to evaluate two different algorithms used to generate the filtering models: SVM (Support Vector Machine) with linear kernel and RF (Random Forest). Additionally, we use nine configurations for training and test sets, varying the training and test percentages from 90-10 to 10-90, we performed 5-fold cross-validation [2], and we report effectiveness in terms of accuracy, i.e., the percentage of true positives for all positive predictions. Moreover, we evaluate four different sources of textual features extracted from the business pages: i) Title (TO); ii) Title (TD) + Description; iii) Title + Full Text (TF); iv) Title + Description + Full Text (ALL).

Table I shows the accuracy of each leaning algorithm used to filter real estate development news with different features for each training and test configuration schema. From Table I we observe that the title-only feature is less effective than the others, since it contains few words that are related to the business context. Additionally, we observe that SVM mostly outperforms RF with accuracy from 92% to 100% depending on the number of instances used in training.

TABLE I
Filtering real estate development news Accuracy.

Config.	RF				SVM			
	TO	TD	TF	ALL	TO	TD	TF	ALL
90/10	0,70	**1,00**	0,90	0,93	0,80	0,96	0,96	**1,00**
80/20	0,65	0,96	0,83	0,90	0,73	0,96	**0,98**	**0,98**
70/30	0,64	0,90	0,93	0,86	0,71	0,89	**0,97**	0,92
60/40	0,68	0,81	0,88	0,95	0,73	0,91	**0,97**	0,91
50/50	0,71	0,86	0,87	0,90	0,75	**0,95**	0,87	0,93
40/60	0,70	0,82	0,77	0,83	0,73	0,87	0,86	**0,92**
30/70	0,72	0,72	0,71	0,88	0,71	0,86	**0,93**	0,91
20/80	0,71	0,52	0,72	**0,92**	0,72	0,75	0,90	0,90
10/90	0,72	0,71	0,71	0,78	0,73	0,79	0,87	**0,92**

Recalling our first and second research questions, these observations attest the effectiveness of our approach to collect and filter real estate development news. In addition, we show that our textual features provide impact positively in the filtering performance.

*This work is funded by CEMIG ANEEL R&D Project GT641 (Brazil).

References

[1] F. Hamborg, N. Meuschke, C. Breitinger, and B. Gipp, "news-please: A generic news crawler and extractor," in *Proceedings of the 15th International Symposium of Information Science*, 2017.

[2] R. Jain, *The art of computer systems performance analysis: Techniques for experimental design, measurement, simulation, and modeling*. Wiley-Interscience, New York, 1991.