Celebrity Watch: Browsing News Content by Exploiting Social Intelligence

Omar Ali, Ilias Flaounas, Tijl De Bie, and Nello Cristianini

Intelligent Systems Laboratory, Bristol University,
Merchant Venturers Building, Woodland Road, Bristol, BS8 1UB, United Kingdom
{Omar.Ali,Ilias.Flaounas,Tijl.DeBie,Nello.Cristianini}@bristol.ac.uk
http://patterns.enm.bris.ac.uk

Abstract. Celebrity Watch is an automatically-generated website that presents up-to-date entertainment news from around the world. It demonstrates the application of many pattern analysis methods that allow us to autonomously monitor millions of news articles and hundreds of millions of references to people mentioned in them. We apply statistical methods to merge references into people, track their association to various topics of news, and generate social networks of their co-occurrences in articles. From this sea of data we select the forty most-relevant people and display them on the website, offering users a highly condensed view of the latest in entertainment news. The site updates itself throughout the day and is the final step in a large, fully-autonomous system that monitors online news media.

Keywords: news mining, statistical inference, trend detection, social networks, entertainment news.

1 Introduction

In this paper we present Celebrity Watch, which is a website that delivers the latest entertainment news from the perspective of the people who appear in it. It is an entirely automated system that updates itself many times per day and is able to detect people who are currently 'trending'.[1] Figure 1 shows the website, which is accessible online at http://celebwatch.enm.bris.ac.uk/.

The website has sections dedicated to those people who are trending today, as well as to those who are trending this month. In the first case we see people who are mentioned in breaking entertainment news, while in the second we see the most popular celebrities of the moment.

Associated with each person are a selection of timeline views depicting their media activity in recent history. Each person also has an automatically inferred social network, which allows users to browse those who are most connected to them, as seen in online news.

[1] The term 'trending' is taken from Twitter (http://www.twitter.com), and refers to a sudden increase in usage of a term on their website.

D. Gunopulos et al. (Eds.): ECML PKDD 2011, Part III, LNAI 6913, pp. 613–616, 2011.

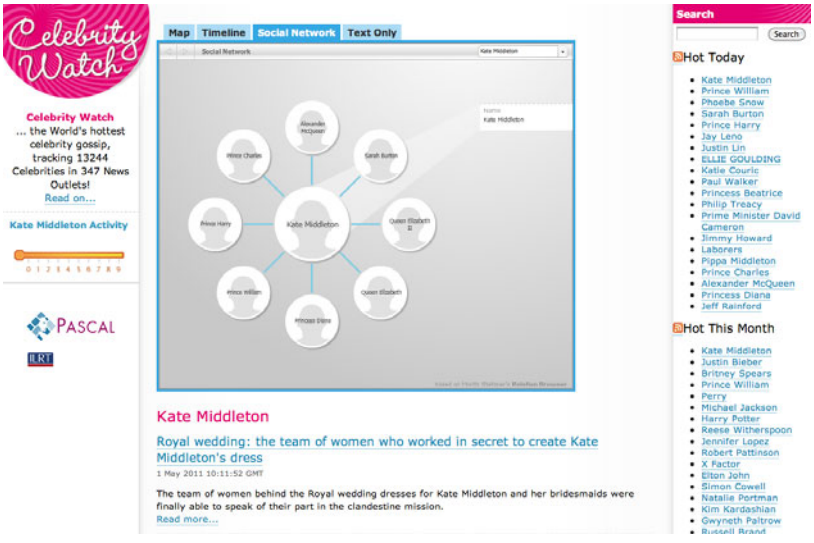

Fig. 1. An image of the main page of Celebrity Watch, centred on Kate Middleton's social network (taken on 2nd May 2011)

Celebrity Watch is the result of an autonomous software pipeline that collects articles from over one thousand online news outlets. The pipeline extracts and resolves references to people mentioned in these articles using statistical methods and multiple sources of information. Our system currently monitors over 15 million English-language articles and tracks over 36 thousand people. A key challenge of building such a system is extracting a meaningful signal of sufficient quality from a large volume of data, while allowing fast and easy access to the latest changes in the world.

News articles are collected for a whole range of topics, and their statistical association to each person is automatically monitored. This allows us to detect interesting people—those whose association to entertainment news is suddenly increasing—and report them to the world immediately. Note that 'entertainment' is just a parameter in the construction of this website; in actual fact we could change this to 'sport' or 'business' to generate a similar site with a different focus.

We should reiterate that the entire pipeline is automated. It collects and labels articles by their topic, extracts and resolves references to people, monitors their association to entertainment news, detects interesting changes in their association, generates social networks, and updates itself throughout the day. All of this without human intervention.

This paper briefly outlines the components of our software pipeline that contribute to the operation of Celebrity Watch. It is updated at regular intervals every day and can be accessed online at `http://celebwatch.enm.bris.ac.uk/`.

2 Methods

Celebrity Watch is generated as a result of many software modules that interact via a number of databases. Online news websites are monitored by a multi-threaded spider, which downloads and parses news feeds in order to populate a database of articles. Much of our data-collection system is already described elsewhere [5], so we omit full details here.

References to named entities are extracted using GATE [3], prior to applying large-scale entity matching using multiple sources of information [1].

The result of these steps is a set of over 36 thousand people linked to over 15 million articles, all of which is automatically processed. In the following section we outline how we infer the topic of a person and how we track their association to entertainment news such that we can immediately detect changes in it. We then explain how social networks are generated, before briefly explaining how we rank articles and geo-locate them.

Topic Tracking: We track the statistical association between named entities the topics of news with which they are associated. This is measured using the odds ratio [2]. All entity–topic associations are tracked using a set of exponentially weighted moving averages. Each of these decays with half-lives of one day, one week, one month, or one year; the fast-decaying averages reflect media activity of a person at the present moment, while the slow-decay ones reflect long-term trends in any media topic. Large increases in association of a person to any topic is typically evidence of a new story about them. We detect such increases, or trends, by *comparing* moving averages.

Celebrity Watch is generated many times each day and its focus is dedicated to only 40 people: the top 20 movers today as compared to this week, and the top 20 movers this month as compared to this year. Our infrastructure allows us to generate a list of the biggest movers within any topic of news, within a few seconds. The remaining steps collect other information from our systems to bring this information together to form Celebrity Watch.

Social Networks: Social networks are generated based on co-occurrences between people seen in entertainment articles. We first construct a master network that considers co-occurrences between people mentioned in entertainment articles seen in the past 30 days. This network is filtered using the χ^2 test of independence, so that only the most significant connections are maintained. This step takes under 10 minutes and is run once per day.

We further filter this network to include only the 40 people of interest at present, as well as those who are within two steps of them in the network. This step takes a few seconds and ensures that the site remains focused on the most interesting people of the moment.

Assembly and Presentation: Final steps in the production of Celebrity Watch include the addition of the most recent articles that mention each person. These are ranked according to where in the article a person is mentioned. Only the

title and summary of each article is displayed, along with a link to the source web page. In addition, we use extracted locations from the text of each article to geo-locate them, to display on the site map.

3 Discussion and Conclusions

The barrage of information delivered on the web is unlikely to slow down, so we need ways to reduce this data overload and focus our attention on that which is most interesting to us. In this case we have presented a digest of the most interesting news by presenting a handful of people of interest, discovered among millions of possibilities.

It demonstrates what can be achieved by integrating multiple machine learning and text mining technologies into a unified, autonomous system. Our approach contrasts with existing systems, [4], [6], in that it presents news stories by detecting the *people* who are most interesting in the world at present.

Further steps could leverage our natural tendency to process emotions associated to news and social relations. We intend to add to the social aspect of the site, so that users may browse the news from a social perspective, and better map to our socially-orientated brains.

Acknowledgements. We would like to thank Simon Price and Ben Joyner of the University of Bristol Institute for Learning and Research Technology, for their work on the web site front-end.

References

1. Ali, O., Cristianini, N.: Information Fusion for Entity Matching in Unstructured Data. In: Papadopoulos, H., Andreou, A.S., Bramer, M. (eds.) AIAI 2010. IFIP Advances in Information and Communication Technology, vol. 339, pp. 162–169. Springer, Heidelberg (2010)
2. Boslaugh, S., Watters, P.A.: Statistics in a Nutshell: A Desktop Quick Reference (In a Nutshell (O'Reilly)). O'Reilly Media, Sebastopol (2008)
3. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: GATE: A framework and graphical development environment for robust NLP tools and applications. In: Proc. of the 40th Anniversary Meeting of the Association for Computational Linguistics, pp. 168–175 (2002)
4. Das, A.S., Datar, M., Garg, A., Rajaram, S.: Google news personalization: scalable online collaborative filtering. In: Proceedings of the 16th International Conference on World Wide Web, WWW 2007, ACM, New York (2007)
5. Flaounas, I., Ali, O., Turchi, M., Snowsill, T., Nicart, F., De Bie, T., Cristianini, N.: NOAM: News Outlets Analysis and Monitoring System. In: SIGMOD (2011), accepted for publication
6. Pouliquen, B., Steinberger, R., Deguernel, O.: Story tracking: linking similar news over time and across languages. In: Proceedings of the Workshop on Multi-source Multilingual Information Extraction and Summarization, MMIES 2008, pp. 49–56. Association for Computational Linguistics, Stroudsburg (2008)