The Migraine Radar - A Medical Study Analyzing Twitter Messages?

Dirk Reinel, Sven Rill, Jörg Scheidt, Florian Wogenstein
Institute of Information Systems (iisys)
University of Applied Sciences Hof
95028 Hof, Germany
{dreinel,srill,jscheidt,fwogenstein}@iisys.de

***Abstract*—This paper discusses the work in progress of the "Migraine Radar" project. The purpose of the project is to validate or disprove the assumed correlation between migraine attacks and weather conditions, especially weather changes. There have been various medical studies on this topic, but the correlation could not be proved with sufficient statistical significance so far. Furthermore, the results of some of the studies are contradictory. For this study, data from the micro-blogging platform Twitter will be analyzed. Twitter messages ("tweets") announcing currently or recently happened migraine attacks are retrieved using the Twitter API (Search-API, REST-API - Standard APIs provided by Twitter to retrieve tweet and user data). Weather data from weather information services are linked to the tweets, using the location information from Twitter. For the German language area, the results will be compared with the results obtained from a set of migraine announcements collected with the help of a web form in the same period of time. First statistics indicate that the number of migraine attacks announced in Twitter exceeds the number of cases in former classical studies by far. The project offers a wide range of possibilities to analyze Twitter messages with regard to migraine attacks. Beside the main purpose, it is also possible to analyze the distribution of migraine attacks over the weekdays or over the seasons. Furthermore an investigation of the spatial distribution of migraine attacks is possible. Instead of weather data, other information can be linked to the migraine sample as well. One example could be air pollution data.**

***Index Terms*—migraine; trigger; Twitter; weather; text mining**

I. Introduction

About 10% of the population suffer from migraine. Several factors are assumed to trigger migraine attacks. Examples for these potential trigger factors are food, stress, hormonal disbalance and sleep irregularities [1]. Especially weather conditions or changes in weather conditions are supposed to cause migraine attacks, although a clear correlation could not be proved in several studies. Typical problems of traditional studies in this area are that only a small number of patients were involved and only patients coming from a small local area were considered. Furthermore some studies were restricted to a short period of time or only took very severe cases of migraine attacks [2] into account.

Using data from Web 2.0 platforms like Twitter perhaps can help to solve some of these problems, but it must be admitted that other difficulties are expected to occur. For example, it could be a problem that the data will contain more noise insofar as in some cases there will be ambiguities while identifying the migraine tweets, e.g., it might be difficult to separate migraine and normal headache cases.

Thus, beside the main purpose of the study, another aim is to find out whether it is possible to use data from social networks to carry out or support medical studies and to identify typical problems while doing so.

In Section II, some related work is introduced. Section III contains a brief description of the Migraine Radar project and explains the main input data and how it is organized for further analysis. In Section IV, the data analysis approach is discussed. A short conclusion completes the paper in Section V.

II. Related Work

Several studies concerning the trigger factors of migraine attacks have been performed over the last decades. As weather conditions and changes in weather conditions are among the factors mentioned most frequently when discussing possible trigger factors, many studies have set their focus on the investigation of this correlation.

A former study investigated the correlation between weather changes and both tension headache and migraine [3]. Another purpose of this study was the investigation of the weekday dependence of migraine attacks. Weather data from a local Institute of Meteorology was used to enrich the clinical data of the migraine patients. The authors found evidence of a weekday dependence. They also saw a correlation to a change in atmospheric pressure 1-3 days after the attack. Limitations were the small number of patients under investigation and the restriction to a small local area.

Another study to investigate many possible trigger factors was performed using patient data from a headache clinic [1]. A detailed headache evaluation was performed, but without adding weather information to the data. Concerning the correlation between weather and migraine attacks, the author found out that about 50% of the patients report attacks occasionally triggered by weather changes.

The approach of Prince et al. [4] was somehow different. The authors evaluated headache calendars provided by 77 migraine patients in a headache clinic. They first asked the patients whether they believe that weather is a trigger for migraine attacks for them. Afterwards, they linked weather

 ISBN: 978-1-61208-162-5

data from the the National Weather Service to the data. Three weather factors were considered. The result was, that about 50% of the patients were found to be sensitive to at least one weather factor, more patients thought they were sensitive but were not.

A lot of research work on the analysis of Web 2.0 platforms is currently going on. Twitter as a micro-blogging platform offers some advantages compared to others. It provides an almost real-time access to utterances of user's daily life as well as insights into their thoughts, opinions and sentiments. In [5], Bifet and Frank give a brief introduction to Twitter as a micro-blogging platform. They also discuss the access to data using the Twitter API and some possibilities in the area of sentiment knowledge discovery using Twitter. In [6], the authors discuss the possibillity of monitoring earthquake occurencies using data from Twitter.

First studies in the medical sector have already been carried out using Twitter messages. In [7], for example, the author analyzed 500 million tweets to investigate an influenza outbreak in the United States enabling him to forecast future influenza rates with high accuracy. In [8], the authors present an automated tool for tracking the prevalence of influenza-like illness using Twitter messages.

Also Google search query data was already used to track influenza epidemics; see [9].

III. The Migraine Radar

A. General Overview

The project "Migraine Radar" collects announcements of migraine attacks from Twitter messages and from a web form. The data are enriched with weather data. Afterwards, the data are analyzed in order to investigate the correlation between the number of evident migraine attacks and certain weather conditions or changes to the weather conditions.

B. Input Data

Several data sources are important in the Migraine Radar project:

- Short messages announcing migraine attacks in the micro-blogging platform Twitter are retrieved with a simple search for "migraine" / "migräne" for the English / the German language. Due to the fact that the Twitter limit of 150 requests per hour is not reached, all relevant tweets are recorded. In some cases the spatial location of the patient is part of the tweet (about 2% of the tweets), in 80% of the cases it can be retrieved from the position the Twitter user has stored in his profile. If no position is available, the tweets will be discarded. Usernames are eliminated in order to anonymize the tweets. For normalization purposes (see section IV-C), tweets containing other search terms are also retrieved.
- Migraine attacks can also be reported using a web form [10], which is available only for Germany at the moment. In this form patients can add some additional information such as age, gender and severity of the migraine attack, making some additional analyses possible later on.
- In order to link the location of the tweets or the entries in the web form to an actual city or town, spatial information provided by GeoNames [11] is used.
- Historical and latest weather data are taken from weather information systems such as the "Deutscher Wetterdienst" [12] for Germany or the Weather Underground [13] for the United States.

C. Data Organization

All raw data are stored in a database. As soon as the preprocessing and the preselection of the migraine announcements is finished, the data are organized in a multidimensional data model to provide an easy and fast access for further analysis.

IV. Data Analysis

The analysis of the data will be performed in several steps. Firstly, a clean sample of migraine attack announcements with reliable spatial information and a defined start day of the attack has to be obtained (see chapters IV-A and IV-B). Afterwards, correlations between weather variables (e.g., temperature, pressure, wind, etc.) or their changes and the number of migraine attacks will be analyzed. Therefore, a normalization of the tweets is necessary (see chapter IV-C). Finally an investigation of the systematic uncertainties of the study has to be carried out (see IV-E).

A. Preprocessing and Preselection of the Tweets

Typical examples of tweets retrieved are as follows:

- *"UGH MIGRAINE"*
- *"Last Saturday, I woke up with an unbelievable migraine!"*
- *"Migraine solution for fast relief <some URL>"*
- *"@<some user> Maybe you're getting a migraine!"*
- *"Migraine, why aren't you going away? It hurts!"*

There are two important filtering steps required in order to select the tweets relevant for analysis:

1) Advertising: Tweets with recommendations, for example for medicine for migraine treatment, are normally not announcing an actual migraine attack. In most cases, they provide a web link to the advertised medicine. In order to eliminate tweets containing advertising, all messages that include web links are discarded.
2) Answering tweets: In most cases answers to other tweets are no real migraine attack announcements. Accordingly all tweets with "@"-signs have to be discarded.

A manual investigation of a part of the tweets indicates that more than 90% of the filtered tweets are announcements of migraine attacks and that the acceptance of good tweets is not affected too much by these steps. Precision and recall will be calculated during the final analysis to confirm this assumption.

B. Selection and Evaluation of the Tweets

The tweets remaining after the preselection steps have to be regarded in more detail. For each tweet it has to be decided:

1) Whether the tweet really refers to a migraine attack of the Twitter user writing the message.

 ISBN: 978-1-61208-162-5

2) At which time the migraine attack started. The first day of the attack is the relevant information, a finer resolution in time will not be achievable in the majority of cases. Some tweets announce actual attacks (like *"UGH MIGRAINE"*), some indicate a defined start day in the past (*"Yesterday I got a horrible MIGRAINE"*). But also ambiguous cases are occurring. As migraine attacks sometimes last up to three days, in the last example from above (*"Migraine, why aren't you going away? It hurts!"*), it cannot be decided when the attack started. Consequently tweets like this have to be discarded as well.

In order to perform this step of the analysis, text mining techniques have to be applied. One possible approach is to identify patterns which indicate an actual migraine attack (e.g., *"I ... have/had ... migraine"*) or to identify typical words occurring in migraine tweets (a first look at the remaining tweets shows that announcements of migraine attacks often contain swearwords).

Also some special cases have to be considered (e.g., *"migraine"* in names, in song titels etc.).

C. Normalization

In order to find out whether an increase in the number of migraine tweets really indicates a rise of migraine attacks under certain weather conditions or whether it is due to a systematic fluctuation, e.g., a general rise of Twitter activities, the tweets have to be normalized. As the total number of tweets for a specific date and place is not available, the normalization has to be obtained from tweets retrieved by searching for some other keywords for each of the two languages. For the English language, such words are for example "coffee", "time", "money", "nature" and "day". For the German language, the translated keywords ("Kaffee", "Zeit", "Geld", "Natur", "Tag") are used. In order to keep the number of normalization tweets small, individual downscaling factors are adjusted in a way that the number of downscaling tweets still exceeds the one of the migraine tweets.

The preselection criteria used for the migraine tweets are also applied to the normalization tweets. Especially tweets with web links (often containing advertising) and answers to other users (detected by the presence of "@"-signs) will be discarded.

D. Preliminary Data Collection Statistics

Table I summarizes the numbers of tweets retrieved per language for a period of four weeks. It also shows the effect of the preprocessing steps (elimination of advertisements and answers to other Twitter users). A first look at the remaining tweets shows that about 60% of them are tweets really announcing a migraine attack. This allows a rough estimation of the total number of migraine attacks recorded in a period of one year. A much more detailed analysis of the effect of the preprocessing steps remains to be done within the final analysis. Also the quality of the geo-location data, depending on whether the information is available directly from the tweet or only from the user profile, has to be analyzed.

Selection Step	**Number of Tweets**		
	English	**German**	**Sum**
Raw Data	81,444	1,057	82,501
Position Available	65,477	819	66,296
Without Links	56,355	600	56,955
Without @-signs	39,082	333	39,415
Exp. migraine tweets (4w)	$\approx$ 24,000	$\approx$ 200	
Exp. migraine tweets (1y)	$\approx$ 300,000	$\approx$ 2,400	

TABLE I
DATA COLLECTION STATISTICS FOR A FOUR WEEK CRAWLING PERIOD

E. Additional Remarks on the Data Analysis

Several follow-up studies have to be conducted in order to evaluate the systematic uncertainties of a study based on messages from Twitter:

- All the selection steps, including the preselection, have to be reviewed manually and performance indicators (e.g., precision, recall) have to be calculated.
- For the data sample in German, results obtained using the Twitter messages have to be compared with the results based on the data collected via the web form. As the information content of these data is richer and more reliable many systematic checks are possible. For example, the web form data contains a subset of messages (entered as migraine attacks with aura) where the probability of real migraine attacks is very high.
- The normalization can be the crucial point in the analysis. It has to be investigated, for example, the error of an improper normalization caused by different probabilities of normalization tweets during the day (e.g., there will be more *"coffee"* tweets in the morning).

F. Possibilities for Further Studies Using the Collected Data

The data collected for this study offer numerous further possibilities for doing research. Examples for other opportunities are:

- Investigation of the weekday dependence of migraine attacks: it is stated by many sources that migraine attacks occur more frequently on Saturdays or Sundays, because some of the attacks are triggered by decreasing stress. Several studies, e.g., Osterman et al. [3] and Morrison [14] investigated the weekday dependence of migraine attacks. They came to different results and the statistical significance was low. Similarly, season dependency of migraine attacks can be analyzed.
- Some migraine tweets contain information which allows to determine the starting time of the migraine attack (for example *"Migraine since three hours ..."*). Filtering these migraine tweets allows a study of the daytime dependence of migraine attacks with a chance to reach a higher statistical significance than former studies.

 ISBN: 978-1-61208-162-5

- The regional distribution of migraine attacks (e.g., urban vs. rural regions) can be examined.

V. CONCLUSION

Modern Social Media platforms like Twitter or Facebook offer enormous possibilities for information retrieval using modern data analysis techniques like data- and text mining.

The purpose of this study is to investigate the possibilities of conducting medical studies using data from Web 2.0 platforms. A first look at the data collected from Twitter seems to be promising, especially the number of migraine attacks analyzed is much higher than in classical clinical studies. But, of course, the approach discussed in this paper suffers from several limitations. In a classic clinical study, doctors diagnose migraine and ascertain information concerning the migraine. Using data from Web 2.0 platforms, it is not possible to verify that migraine announcements are really made by migraine patients.

First results of the project are expected by the end of 2011.

ACKNOWLEDGMENTS

The authors would like to thank all members of the Institute of Information Systems (iisys) in Hof for many helpful discussions and especially R. Göbel for his great efforts on foundation of the institute.

This work is supported by the German Federal Ministry of Education and Research (BMBF). The Institute of Information Systems is supported by the Foundation of Upper Franconia and by the State of Bavaria.

REFERENCES

[1] L. Kelman, *The triggers or precipitants of the acute migraine attack*, Cephalalgia, 2007, **27**, pp. 394-402.

[2] W. J. Becker, *Weather and migraine: Can so many patients be wrong?*, Cephalalgia, 2011, **4**, pp. 387-390.

[3] P. O. Osterman, K. G. Lövstrand, P. O. Lundberg, S. Lundquist, and C. Muhr, *Weekly Headache Periodicity and the Effect of Weather Changes on Headache*, Int. J. Biometeor. 1981, vol. 25, number 1, pp. 39-45.

[4] P. B. Prince, A. M. Rapoport, F. D. Sheftell, S. J. Tepper, and M. E. Bigal, *The Effect of Weather on Headache*, Headache, 2004 Jun; 44(6), pp. 596-602.

[5] A. Bifet and E. Frank, *Sentiment knowledge discovery in Twitter streaming data*, Proceedings of the 13th international conference on Discovery science (DS'10), B. Pfahringer, G. Holmes, A. Hoffmann (Eds.) 2010, Springer-Verlag Berlin, Heidelberg, pp. 1-15.

[6] T. Sakaki, M. Okazaki, and Y. Matsuo, *Earthquake shakes Twitter users: real-time event detection by social sensors*, Proceedings of the 19th international conference on World wide web (WWW '10), 2010, ACM, New York, pp. 851-860.

[7] A. Culotta, *Detecting influenza outbreakes by analyzing Twitter messages*, www2.selu.edu/Academics/Faculty/aculotta/pubs/culotta10detecting.pdf, (accessed July 26, 2011).

[8] Vasileios Lampos, Tijl De Bie, and Nello Cristianini, *Flu Detector - Tracking Epidemics on Twitter*, European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD) 2010, Springer-Verlag Berlin, Heidelberg, pp. 599-602.

[9] J. Ginsberg, M. Mohebbi, R. Patel, L. Brammer, M. Smolinski, and L. Brilliant, *Detecting influenza epidemics using search engine query data*, Nature **457**, 2009, pp. 1012-1014.

[10] Institute of Information Systems, *The migraine radar*, www.migraene-radar.de (accessed July 26, 2011).

[11] Marc Wick (Founder), *GeoNames*, www.geonames.org (accessed July 26, 2011).

[12] Das Bundesministerium für Verkehr, Bau und Stadtentwicklung, *Deutscher Wetterdienst*, www.dwd.de (accessed July 26, 2011).

[13] Weather Underground Inc., *Weather Underground*, www.wunderground.com (accessed July 26, 2011).

[14] Morrison, *Occupational stress in migraine - is weekend headache a myth or reality?*, Cephalalgia, 1990, **10**, pp. 189-193.

 ISBN: 978-1-61208-162-5