Enhancing the Curation of Botanical Data Using Text Analysis Tools

Clare Llewellyn[1], Clare Grover[1], Jon Oberlander[1], and Elspeth Haston[2]

[1] University of Edinburgh, Edinburgh, United Kingdom
C.A.Llewellyn@sms.ed.ac.uk,
{Grover,Jon}@inf.ed.ac.uk
[2] Royal Botanic Gardens Edinburgh, Edinburgh, United Kingdom
E.Haston@rbge.ac.uk

Abstract. Automatic text analysis tools have significant potential to improve the productivity of those who organise large collections of data. However, to be effective, they have to be both technically efficient and provide a productive interaction with the user. Geographic referencing of historical botanical data is difficult, time consuming and relies heavily on the expertise of the curators. Botanical specimens that have poor quality labelling are often disregarded and the information is lost. This work highlights how the use of automated analysis methods can be used to assist in the curation of a botanical specimen library.

Keywords: text analysis, text mining, geographical location, assisted curation, botany.

1 Introduction

The aim of this work is to improve the interaction between users and automatic text analysis tools within a practical context. A tool has been created that allows users to curate botanical data by adding geographical locations. This is achieved via the user interacting with automatically generated locations extracted from textual records about botanical specimens. The user can correct or make additions to this generated output in order to specify the exact location where the botanical sample was collected.

To pursue the aim of creating a productive usable interface for textual analysis tools, a specific interface for such a tool has been created. The tool and interface are both applicable to many uses, but in order to evaluate it in detail, the focus is botanical science specimen data. The user group that this tool has been tested by are staff the Royal Botanic Gardens, Edinburgh (RBGE). The data curation experts at the RBGE expressed a need for the integration of a tool that extracts geographical locations from plant specimen data records into their current work flow. The demand for such a tool gives the opportunity to engage its likely users in evaluating the interface, therefore producing valid usability results.

P. Zaphiris et al. (Eds.): TPDL 2012, LNCS 7489, pp. 480–485, 2012.

2 Background

The Royal Botanic Garden of Edinburgh an internationally renowned centre of excellence for plant biodiversity research. The herbarium houses nearly three million specimens representing half to two thirds of the world's flora. It holds specimens collected from across the world and continues to receive approximately ten thousand new specimens each year [13].

Plant specimens are labelled with data that is relevant from their collection. Geographic referencing in this domain means converting the textual descriptions of where a plant was collected into machine readable geographic locations generally using a map based coordinate system. This is either done at the time when the plant is collected by GPS systems or retrofitted from textual descriptions [2, 9]. Historically, locations on plant specimens have been vague. Identifying and correcting plant specimens records that contain errors is time consuming and expensive for curators [9], therefore improvements to the speed and the accuracy of this process would be valuable. Currently, geographic referencing is conducted manually using resources such as gazetteers and maps to find the coordinates of the place names that have been identified in the plant specimen records by the curators. Tools have been built to assist in this process, including BioGeomancer [14] and GEOLocate [15], but these systems do not always fit well into the curation workflow [9]. Once the data has been geographically located it allows a botanical scientist to study environmental changes, particularly those concerning human impact and climate change.

Locations can be automatically generated through content analysis, natural language processing and text mining [3]. Widespread use of text analysis has not yet been achieved. For example in geographic referencing it is believed that the main barrier to uptake is that accuracy levels usually fall short of the expectations and needs of the user. It is proposed that this problem is rectifiable through the provision of interface extensions to existing text analysis tools to allow the user to correct and enhance automatically created output, thereby combining the efficiency of automatic processing with the accuracy of manual annotation [1]. Metrics for text analysis evaluation currently focus on comparisons with other text mining systems rather than evaluating the usefulness of the tool within a domain [1,10,12]. A study by Alex et al. in 2008 [1] found that the speed of curation can be increased by a third by assistance of text mining tools.

3 Prototype Tool

3.1 Data

Plant specimens have labels describing the collection details of that specimen. The text from these labels are stored in a database. The conversion of the label to a record is a manual process performed by the curators. This study focused on records from the United Kingdom and Ireland from 1747 to 2010. The total number of records processed was 43,060. It was found that, in total, 63.82% of records had some degree of geographical information, and could be geolocated.

3.2 Text Analysis Tools

The data from the database was processed using the Edinburgh Informatics information extraction tools which include LT-TTT2 and the Edinburgh Geoparser [6,7,11]. These are well established tools that process text and XML to identify place names and provide geographic coordinates for the locations. The Geoparser is made up of two main components; the Geotagger which provides place name recognition (identifies text strings as places) and the Georesolver which provides geographic referencing (looks up the names in a geographic gazetteer) [6]. The named entity recognition tool identifies word sequences as place-name entities and marks them up as XML elements. After initial tokenisation and part-of-speech tagging, it uses a rule-based method that takes into account information about part-of-speech, capitalisation, local context and lexicon look-up. The place-name entities recognised by this method are converted to gazetteer queries which are submitted to one of the Unlock or GeoNames gazetteer services [6,7,11].

3.3 The Tool

The data curation experts requested an automatic tool that extract geographical locations from plant specimen data records and could be integrated into their current work flow. The data was initially processed through text mining, database matching of similar fields and a National Grid Reference conversion to latitude and longitude. The information produced from this processing was stored in the database. The system is web based and the users interact with it through a webserver to query the database. The interface provides two views of the result, as a list of locations and as points on a map. The maps used are accessed through APIs - Google Maps and the National Library of Scotland's Ordinance Survey Maps.

4 Evaluation

An evaluation was conducted to test the hypothesis that a textual analysis tool can be used to improve, increase the speed or accuracy of the workflow of curators who are archiving plant specimen data. The current manual curation process was compared against a tool with textual analysis support. The evaluation was conducted in a manner adapted from a digital library evaluation framework [5]. Human computer interaction (HCI) within digital libraries has been studied extensively for the past ten years [8]. The tool was evaluated to ensure that it observed the basic HCI principles of a digital library such as obtaining correct results to a query quickly (precision and recall)[4]. It is important for the user to receive a manageable number of results so that they can see what the general content will be. Furh et al (2007)[5] provide an extensive framework for the evaluation of digital libraries which is adapted for this task. They suggest focusing on usability, usefulness (or relevance) and performance; therefore these were the areas evaluated in this work.

The evaluation was conducted with ten participants all of whom work at the RBGE. Each participant was asked to perform eight tasks. The data used for evaluation was data with known locations (a random sample from the 881 RBGE records that contained latitude and longitude values). This was then used to provide an accuracy measure for each task. A post task interview was used to provide qualitative information on the participants' opinion of the system.

Usability is measured by looking at the effectiveness, adaptability, enjoyability and learnability of the tool. Effectiveness was measured by how many tasks could be completed [4,5]. The results suggest that the tool performs slightly better than the traditional method. Adaptability was measured by whether they could adapt the experience to their own preferences. Comments on these features suggested that they were generally well liked. Enjoyability and learnability were measured through satisfaction scores for ease of use, visual appearance, contents, structure, error corrections and usefulness of help information. The tool scored highly in this category, the users liked the tool and found it easy to use. Participants found the tool easier to use than the traditional method.

Usefulness is evaluated through the relevance of provided content. If the content assist with the task defined in a satisfactory way [4,5] and whether the content provided led to participants accurately locating the samples. The text mining suggestions were not considered completely accurate, as many false positives were returned in order to include as many true positives as possible, but the text mining suggestions were still considered helpful. The users were willing to tolerate a degree of inaccuracy in the suggestions. Initially it was found that there was no significant difference between the accuracy of the two systems. However, there is a significant positive correlation between the tasks showing that some task was difficult with either method.

In order to look more closely at the accuracy achieved using the tool a further experiment was conducted. All participants in the test were asked to geo-locate all of the samples used in the initial evaluation using the tool. These locations were then clustered and the location which was the furthest from the others was left out, as was any location more than 25km away from the average point. Using an average location point from those left, a significant increase in accuracy was found when using the tool ($p= 0.012$, $t=-2.742$, $df= 23$). Thus the tasks may be difficult for specific individuals but when an average is taken over the whole group the result will be accurate.

Performance and efficiency of the tool was evaluated by assessing the efficient retrieval of information. It was measured by how much time it took to correctly complete tasks [4,5]. Each task was timed for each participant. In addition participants were asked to rate performance for each task. The performance, which was judged by speed of task completion, was better on average with the tool (see table 1). A paired sample t-test shows that the difference is not significant ($p=0.539$, $t=-0.639$ $df=9$).

Table 1. Average Speed for Tasks (in seconds)

	Minimum Time	Maximum Time	Mean Time	Standard Deviation
Tool	139.25	263.75	190.65	42.47748
Traditional	87.75	300.00	202.93	74.7781

A further experiment was conducted to investigate if the number of text mining locations offered had an effect on the time taken to complete the task to see if there was an optimal number of locations. Initially the total set of locations provided to the user was considered: every single latitude and longitude pair for every place name. Analysis indicates that it is possible that there is a positive effect of either offering very few or very many suggestions (below 2 and above 5). The total number of locations offered was contrasted with the number of unique locations. The tool often suggests a number of individual latitude and longitude locations for a single place name (as many location names are reused). A unique location is classified as a single place name (no matter how many suggestions are offered for that name). It was found that 2 unique locations may be beneficial, possibly because they are used to provide confirmation. With 1 location the task may take longer, as the user may need to consult other features. With more than 3 locations the task may take longer, as the locations may be contradictory.

5 Conclusions

The specific objectives of this work were to identify where text analysis tools can be used in the botanical curation workflow, to design and implement a prototype tool and to evaluate if this tool improves the ease, speed or accuracy of botanical curation. A tool was created that allowed users to interact with automatically generated geographical information in order to correct or make additions to the output. As requested by the data curation experts, the tool has been integrated into the current workflow.

In the evaluation it has been shown that, using the tool, average speeds are quicker. The tool scored highly for both usability and usefulness. The participants in the evaluation liked the tool and found it easy to use - they preferred it to the traditional method of using multiple data sources. The accuracy of the geographic location was compared between the tool and the traditional method. Initially, it was found that there was no significant difference in the accuracy of the two systems. When multiple participants used the tool to identify a location and values were clustered, leaving out the least similar location, a significant increase in accuracy is found. This shows that while the tasks may be difficult for individuals, higher accuracy can be gained from using locations from a number of individuals: they will collectively locate the specimen accurately. This suggests that this is an ideal tool for use with crowd sourcing. The analysis suggests that there may be an optimal number of total text mining suggestions and unique

text mining locations that reduces the burden on the user and leads to more efficient geographic location.

References

1. Alex, B., Grover, C., Haddow, B., Kabadjov, M., Klein, E., Matthews, M., Roebuck, S., Tobin, R., Wang, X.: Assisted curation: does text mining really help? In: Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, pp. 556–567 (2008)
2. Allen, W.H.: The Rise of the Botanical Database. BioScience 43(5), 274–279 (1993)
3. Auer, S., Lehmann, J., Hellmann, S.: LinkedGeoData: Adding a Spatial Dimension to the Web of Data. In: Bernstein, A., Karger, D.R., Heath, T., Feigenbaum, L., Maynard, D., Motta, E., Thirunarayan, K. (eds.) ISWC 2009. LNCS, vol. 5823, pp. 731–746. Springer, Heidelberg (2009)
4. Blandford, A., Buchanan, G.: Usability of digital libraries: a source of creative tensions with technical developments (2003)
5. Fuhr, N., Tsakonas, G., Aalberg, T., Agosti, M., Hansen, P., Kapidakis, S., Klas, C.-P., et al.: Evaluation of digital libraries. International Journal on Digital Libraries 8(1), 21–38 (2007)
6. Grover, C., Givon, S., Tobin, R., Ball, J.: Named Entity Recognition for Digitised Historical Texts. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008) (2008)
7. Grover, C., Tobin, R.: Rule-Based Chunking and Reusability. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy (2006)
8. Jeng, J.: What Is Usability in the Context of the Digital Library and How Can It Be Measured? Information Technology and Libraries 24(2), 46–56 (2005)
9. Johnson, N.F.: Biodiversity Informatics. Annual Review of Entomology 52(1), 421–438 (2007)
10. Dietrich, R.-S., Kirsch, H., Couto, F.: Facts from Text Is Text Mining Ready to Deliver? PLoS Biol. 3(2) (February 15, 2005)
11. Richard, T., Grover, C., Byrne, K., Reid, J., Walsh, J.: Evaluation of georeferencing. In: Proceedings of the 6th Workshop on Geographic Information Retrieval, GIR 2010, 7:17:8. ACM, New York (2010)
12. Winnenburg, R., Wchter, T., Plake, C., Doms, A., Schroeder, M.: Facts from text: can text mining help to scale-up high-quality manual curation of gene products with ontologies? Briefings in Bioinformatics 9(6), 466–478 (2008)
13. Edinburgh Royal Botantic Gardens website, `http://www.rbge.org.uk/`, (accessed March 29, 2012)
14. BioGeomancer, `http://www.biogeomancer.org/`, (accessed August 15, 2011)
15. GeoLocate, `http://www.museum.tulane.edu/geolocate/`, (accessed August 15, 2011)