Classical vs. Crowdsourcing Surveys for Eliciting Geographic Relevance Criteria

Stefano De Sabbata[1], Omar Alonso[2], and Stefano Mizzaro[3]

[1] University of Zurich-Irchel
Winterthurerstrasse 190, CH-8057 Zurich, Switzerland
stefano.desabbata@geo.uzh.ch
[2] Microsoft Corp.
1065 La Avenida, Mountain View CA, USA
omar.alonso@microsoft.com
[3] University of Udine
Via delle Scienze 206, 33100 Udine, Italy
mizzaro@uniud.it

Abstract. Geographic relevance aims to assess the relevance of physical entities (e.g., shops and museums) in geographic space for a mobile user in a given context, thereby shifting the focus from the digital world (the realm of classical information retrieval) to the physical world. We study the elicitation of geographic relevance criteria by means of both a classical survey and an Amazon Mechanical Turk (a crowdsourcing platform) survey. This allows us to obtain three results: first, we gather a set of criteria and their relative importance; second, we gain a first insight on the differences between geographic relevance and classical relevance as commonly understoon in the IR field; and third we draw some considerations on the agreement, on the importance of specific criteria, among the participants to the classical and the crowdsourcing surveys.

Keywords: Relevance, Crowdsourcing, Amazon Mechanical Turk, SurveyMonkey

1 Introduction

The elicitation of relevance criteria dates back to the 90s, if not earlier [7]. Although such criteria seemed quite well established at that time [2], recently this issue is studied again [1]. This is probably due to the Web, that on the one side provides novel search services that might entail a different notion of relevance, and on the other side allows more convenient methods for preparing surveys involving several participants.

In this short paper, we concentrate on Geographic Relevance (GR), a recent area of Information Retrieval (IR), and we discuss the elicitation of relevance criteria by means of:

- SurveyMonkey (SM, www.surveymonkey.com), a Web service that allows the preparation of an online survey whose participants are then invited by email, and
- Amazon Mechanical Turk (AMT, www.mturk.com), a crowdsourcing platform that allows to outsource to the crowd specific tasks for a small amount of money.

The aim of this research is threefold:

- to find suitable GR criteria, that might be different from the classical relevance criteria;
- to gain a first insight into the difference between GR and the classical concept of relevance in the IR field;
- to understand if AMT provides reliable results, or at least if those results agree with the SM ones, which are obtained in a more classical way.

AMT quality and reliability are important issues [6]: there is no guarantee that AMT workers provide reliable answers and that they carry on their task in a reliable way; for example, workers might cheat to quickly gain money. This is even more critical as crowdsourcing is emerging as a widespread alternative for relevance evaluations.

In the following, we first define GR (Section 2) and discuss crowdsourcing and AMT (Section 3) then we present the experimental study and its results (Section 4), and we finally summarize the main findings (Section 5).

2 Geographic Relevance Criteria

The basic idea of GR is to assess the relevance of *physical entities* (e.g., shops and museums) in geographic space for a mobile user in a given context [8]. This definition implies a shift from the informational world — that is the focus of IR, which is devoted to retrieve information from unstructured digital document collections — to the physical world. In other terms, the aim of GR is to apply the principles and concepts developed in the field of IR not only in the informational world, but also in the physical world [3].

GR is different from Geographic Information Retrieval because the second still focuses on digital entities. The aim of Geographic Information Retrieval is to retrieve geographic information from digital documents, or to find relevant digital documents that can satisfy a user's need for geographic information. GR uses digital entities (e.g., the objects in a collection within a Geographic Information System, or documents, or images, etc.) as means to estimate the relevance of the physical entities they refer to, rather than aiming to evaluate the relevance of the digital entities themselves.

In shifting the focus from the digital world to the physical world, a first question is whether the criteria of relevance developed in IR [7, 2, 1] can be applied to assess GR. A second question is whether other criteria are needed in order to fully understand the relevance of a physical entity. We ground our study

Properties	**Geography**	**Information**	**Presentation**
Topicality	Spatial proximity	Specificity	Accessibility
Appropriateness	Temporal proximity	*Availability*	Clarity
Coverage	*Spatio-temporal proximity*	*Accuracy*	Tangibility
Novelty	*Directionality*	*Currency*	*Dynamism*
	Visibility	Reliability	*Presentation quality*
	Hierarchy	Verification	
	Cluster	Affectiveness	
	Co-location	Curiosity	
	Association rule	Familiarity	
		Variety	

Table 1. Four sets of GR criteria, classified as in [4].

on the set of criteria of GR proposed in [4]; these criteria are listed in Table 1. We do not have the space here to discuss these criteria in detail; a comprehensive description of each single criterion, together with a more in depth analysis, is provided in [5].

3 Crowdsourcing

Crowdsourcing has emerged as a feasible alternative for relevance evaluation because it brings the flexibility of the editorial approach at a larger scale.

AMT is an example of a crowdsourcing platform: it is an Internet service that gives developers the ability to include human intelligence as a core component of their applications. Developers use a web services API to submit tasks, approve completed tasks, and incorporate the answers into their software applications. To the application, the transaction looks very much like any remote procedure call: the application sends the request, and the service returns the results. People (the "crowd") come to the web site looking for tasks and receive payment for their completed work. In addition to the API, there is also the option to interact using a dashboard that includes several useful features for prototyping experiments. There is an increased participation by large numbers of online users from all over the world, which is a good sample that includes diversity.

The individual or organization who has work to be performed is known as the *requester*. A person who wants to sign up to perform work is described in the system as a *worker*.

One issue with AMT and similar crowdsourcing platform is quality [6]: there is no guarantee that the workers provide correct answers and that they carry on their task in a reliable way. For example, workers might cheat to quickly gain money. One of the aims of this paper is to compare a survey carried on by means of AMT with a similar one carried on by more classical means, like SM.

1. Considering a place that fits your needs by its category (e.g. a restaurant, if you want to go out for dinner), which other criteria would you take into account?
   - A place that offers **just** the services you need is more relevant than a place that also offers **other** services.
   - A place that offers **all** the services you need is more relevant than a place that offers just **some** of them.
   - A place that was **previously unknown** to you is more relevant than an **already known** place.
2. Considering a place that fits your needs, do you take into account the following criteria related to the presented information and the way it is presented (for example on your mobile device) to judge its relevance?
   - The more information available about a place, the higher is the relevance of the place.
   - The more accurate the information about a place, the higher is the relevance of the place.
   - The more current, recent, timely, up-to-date the information about a place, the higher is the relevance of the place.
   - The more dynamic, active or interactive the presentation of information, the higher is the relevance of the presented place.
   - The more the information about a place is presented in a certain format or style, or offers output in a way that is helpful, desirable, or preferable, the higher is its relevance.

Fig. 1. Questions 1 and 2 as framed in SMs and AMTs1.

4 Experiments

4.1 Experimental design

We selected a subset of the criteria listed in Table 1: the 14 criteria in italics. We chose many of the geographic criteria, leaving out *spatial proximity* and *temporal proximity* (we took into account the *spatio-temporal proximity* that combines both), and *association rule* (which is difficult to explain and can be misunderstood if not explained in detail). We selected two or three criteria from each of the other groups, choosing the easier to explain in a few words and, probably, the most intuitive ones.

Towards the aims stated in Section 1, we ran 3 experiments:

- A SM survey (referred to as SMs) sent by email to researchers and students in IR and similar subjects.
- A first AMT survey (AMTs1) obtained by simplifying the SM survey and by focussing on some items only.
- A second AMT survey (AMTs2) obtained, after the responses to AMTs1, by fine tuning the language to tailor it to the AMT environment, where workers usually are not keen to spend much time on a task.

The questions were asked in an indirect way: for example, we did not ask literally whether "*spatio-temporal proximity* is an important GR criterion"; rather

1. Given a place in the right category (e.g., a restaurant, if you want to go out for dinner), which other criteria would you take into account?
   – A place that offers **just** the services you need is more relevant than a place that also provides **other** services.
   – A place that offers **all** the services you need is more relevant than a place that provides just **some** of them.
   – A place that was **previously unknown** to you is more relevant than an **already known** place.
2. Considering a place that fits your needs, do you take into account the following criteria to judge its relevance?
   – The **more information** available about a place, the higher is the relevance of the place.
   – The more **accurate** the information about a place, the higher is the relevance of the place.
   – The more **current, recent, timely, up-to-date** the information about a place, the higher is the relevance of the place.
   – The more **dynamic, active or interactive the presentation** of information, the higher is the relevance of the presented place.
   – The more the information about a place is presented in a certain **format or style, or offers output in a way that is helpful, desirable, or preferable**, the higher is its relevance.

Fig. 2. Questions 1 and 2 as framed in AMTs2.

we asked whether "it is important to take into account whether the place (or a related event) will be available at the time you will be able to reach it (e.g., whether you can reach the shop before it closes)." The questionnaire included a total of 14 items, arranged into three main questions.

Figure 1 shows two of the three questions (each one grouping some items) as framed in SMs and AMTs1. In SMs, a first page was dedicated to the criteria not related to geographic concepts (e.g., *novelty*), whereas a second page was dedicated to the geography-related criteria. The same items have been used in AMTs1, where the 3 questions were all presented in one page. Figure 2 shows the same items as framed in AMTs2, where we slightly modified the questions (but not the items, that were almost identical to SMs and ATMs1[4]), each one presented in a separate page. Participants assessed each item on a 7-point Likert scale "1 - Strongly disagree" – "7 - Strongly agree" (all the scale values appear on the ordinal axis in Figure 3).

4.2 Results

The number of participants in the three cases is similar: SMs got 53 participants, AMTs1 43, and AMTs2 42 (we discarded two outliers from each AMT survey since they were far too quick). The collected demographics say that participants

[4] The only differences, as shown in the figures, is the change of "offer" into "provide" and the usage of boldface to highlight some terms.

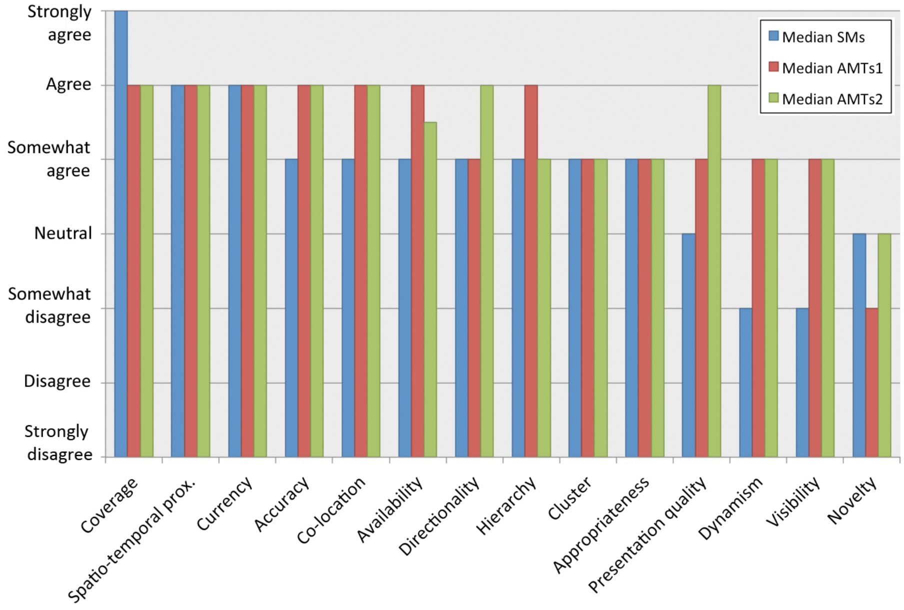

Fig. 3. Median value for each criteria.

to SMs were familiar with digital maps (71% use them at least several times a week), mobile maps (51% use them on their mobile), and online yellow pages (only 30% of the participants have never used them). We did not collect demographic data for AMT (we plan to do that in future experiments). We paid $0.15 to each AMT worker. The total cost for both AMT experiments was $16.

The Kolmogorov-Smirnov normality test was negative, so we considered the variables as ordinal. Figure 3 shows the median importance of the single criteria in the three surveys.

By analyzing the relative importance of the criteria, three groups can be singled out: a first one including the three leftmost criteria (*coverage*, *spatio-temporal proximity*, and *currency*), whose importance seems very high according to all the three surveys; a second group including the central seven criteria whose importance is tangible, but somehow lower with respect to the first group; and a final group of the four rightmost criteria whose importance seems rather low and more inconsistent among the three surveys.

Turning to the agreement among the participants in the three surveys, we can note first that SMs median values are generally lower than AMTs1/2. Also, agreement is different for each criterion, as confirmed by a Mann-Whitney test:

- highly significant ($p < .01$) difference has been found between SMs and AMTs1, and also between SMs and AMTs2, for the criteria *availability*, *accuracy*, *dynamism*, *presentation quality*;

	SM	**AMT**
Demographics	Targeted practitioners and experts	Crowd (unknown workers)
Incentive	Volunteer	Money
Development cost	low	low
Service fee	$30 per month	Free
Participant fee	None	$0.15 per participant
Cost dependencies	Time and service level	Number of participants per survey
Total incurred cost	$60	$8 + $8
Time to completion	45 days	3 days for AMTs1 and 6 days for AMTs2

Table 2. SM vs. AMT comparison.

- highly significant ($p < .01$) difference has been found between SMs and AMTs1 for the criterion *hierarchy*, and between SMs and AMTs2 for the criterion *visibility*;
- significant ($p < .05$) difference has been found between SMs and AMTs1 for the criteria *currency* and *visibility*, and between SMs and AMTs2 for the criterion *co-location*;
- no statistical significant difference has been found between AMTs1 and AMTs2, in any criteria.

Besides differences in quality per se, there are other characteristics that may influence the choice of system for conducting surveys. We present the most important aspects in Table 2.

5 Conclusions

Overall, the results hint that:

- The most important GR criteria seem to be *coverage*, *spatio-temporal proximity*, and *currency*.
- SM and AMT surveys provide slightly different results.
- The differences mainly concern the importance of four criteria (*availability*, *accuracy*, *dynamism* and *presentation quality*)
- None of these four criteria are in the *Geography* set (see Table 1).

This last point is perhaps surprising, since one would expect that the heterogeneous background and cultural differences of the international AMT population would particularly affect the elicitation of geographic criteria. However, in our experiments disagreement was mainly on classical relevance criteria.

One further point to remark is that the average quality of AMT workers answers was good, as demonstrated by the good agreement level with SM, although we did not require qualified workers — as it would have been possible in AMT.

Finally, as future work, we are considering a more “visual” survey, with more images or scenarios, than just pure text as we did in this work .

References

1. O. Alonso and S. Mizzaro. Relevance criteria for e-commerce: a crowdsourcing-based experimental analysis. In *SIGIR '09: Proceedings of the 32nd international ACM SIGIR*, pages 760–761, 2009.
2. C. L. Barry and L. Schamber. Users' criteria for relevance evaluation: A cross-situational comparison. *Information Processing & Management*, 34(2-3):219–236, May 1998.
3. P. Coppola, V. D. Mea, L. D. Gaspero, and S. Mizzaro. The concept of relevance in mobile and ubiquitous information access. In *Mobile HCI Workshop on Mobile and Ubiquitous Information Access*, volume 2954 of *LNCS*, pages 1–10. Springer, 2003.
4. S. De Sabbata. Criteria of geographic relevance. In *6th Int'l Conf. on Geographic Information Science*, 2010.
5. S. De Sabbata and T. Reichenbacher. Criteria of geographic relevance: an experimental study. *International Journal of Geographic Information Science*, forthcoming.
6. P. Marsden. Crowdsourcing. *Contagious Magazine*, 18:24–28, 2009.
7. S. Mizzaro. Relevance: The whole history. *Journal of the American Society for Information Science*, 48(9):810–832, 1997.
8. T. Reichenbacher, P. Crease, and S. De Sabbata. The concept of geographic relevance. In *Proceedings of the 6th Int'l Symposium on LBS & TeleCartography*, 2009.