IPSJ Journal Vol. 52 No. 7 2179–2197 (July 2011)

Regular Paper

Classification of Recommender Expertise in the Wikipedia Recommender System

Christian Damsgaard Jensen,[†1]
Povilas Pilkauskas[†1] and Thomas Lefévre[†1]

The Wikipedia is a web-based encyclopedia, written and edited collaboratively by Internet users. The Wikipedia has an extremely open editorial policy that allows anybody, to create or modify articles. This has promoted a broad and detailed coverage of subjects, but also introduced problems relating to the quality of articles. The Wikipedia Recommender System (WRS) was developed to help users determine the credibility of articles based on feedback from other Wikipedia users. The WRS implements a collaborative filtering system with trust metrics, i.e., it provides a rating of articles which emphasizes feedback from recommenders that the user has agreed with in the past. This exposes the problem that most recommenders are not equally competent in all subject areas. The first WRS prototype did not include an evaluation of the areas of expertise of recommenders, so the trust metric used in the article ratings reflected the average competence of recommenders across all subject areas. We have now developed a new version of the WRS, which evaluates the expertise of recommenders within different subject areas. In order to do this, we need to identify a way to classify the subject area of all the articles in the Wikipedia. In this paper, we examine different ways to classify the subject area of Wikipedia article according to well established knowledge classification schemes. We identify a number of requirements that a classification scheme must meet in order to be useful in the context of the WRS and present an evaluation of four existing knowledge classification schemes with respect to these requirements. This evaluation helped us identify a classification scheme, which we have implemented in the current version of the Wikipedia Recommender System.

1. Introduction

The Wikipedia is a web-based encyclopedia, written and edited collaboratively by Internet users. Over the past decade, the Wikipedia has experienced a dramatic growth in popularity and is by many considered the first source of information on the Internet. The Wikipedia has an extremely open editorial policy that allows everybody to create and modify articles. This has promoted a broad and detailed coverage of subjects[⋆1], but there are plenty of examples of erroneous information that has propagated through the Wikipedia [17),21)]. Providing a means to assess the quality of Wikipedia articles is therefore vitally important for the users to build trust in the Wikipedia and ensure the continued success and growth of the system.

We have identified two different ways to establish trust in content of uncertain provenance, such as articles on the Wikipedia where authors may be anonymous; these methods are *content-based filtering* and *collaborative filtering*. Content-based filtering estimates the quality of Wikipedia articles based on textual properties and the edit revision histories of the article [2),3),6),22),24)]. Collaborative filtering is a filtering technique based on the subjective evaluations, generally called annotations, by other readers [7)], i.e., it uses these annotations to find similar users, then uses the ratings of these similar users to predict future ratings. The Wikipedia Recommender System [11),12),14)], which forms the context of this work, is to the best of our knowledge the only collaborative filtering system for the Wikipedia.

The Wikipedia Recommender System (WRS) was developed to help human users of the Wikipedia to determine the credibility of an article based on feedback from other Wikipedia users. In order to preserve both the large investment that authors have made in terms of time and effort and the familiarity of the user interface for occasional users, the collaborative authoring system must be considered a legacy system that cannot be modified. Moreover, the broad established user base of the Wikipedia means that the WRS should only be offered to users who opt in and must be transparent to everyone else.

The WRS allows users to calculate a personalized rating for any article based on feedback (recommendations) provided by other Wikipedia users. As part of this process, WRS users are expected to provide their own feedback regarding the quality of Wikipedia articles that they have read, so the WRS implements a

†1 Department of Informatics and Mathematical Modelling, Technical University of Denmark

⋆1 At the end of 2010, the English language version of the Wikipedia alone has more than 3.4 million articles [1)].

rating-based collaborative filtering system. The recommendations consists of a simple rating, which encode all quality attributes, such as accuracy, completeness, focus and lack of bias, but also "soft issues" like language and style. This means that it is relevant for all users to provide feedback on all articles that they read, because they may provide useful feedback about the soft issues even if they know little about the subject of the article. Not all recommenders are expected to agree on these attributes, so the WRS implements trust metrics to determine the weight that should be given to the feedback from each individual recommender, i.e., recommendations from recommenders that the user has agreed with in the past will carry more weight in the calculation of the overall rating for the article. The trust metric implemented in the first version of the WRS determines this weight based on all the feedback provided by each individual recommender. However, the scope of the Wikipedia is very broad and recommenders cannot be assumed to be equally knowledgeable in all areas, e.g., some recommenders may provide useful feedback about military history, but may know little about psychology or philosophy. It is therefore important to extend the trust metric, so that it incorporates an assessment of the recommender's expertise in the area of the article. Establishing the areas of expertise for each recommender allows more accurate use of their recommendations when rating the article.

In this paper, we examine the problem of determining the areas of expertise for recommenders in the Wikipedia Recommender System. We do not generally expect recommenders to be known to other users and we do not wish to violate privacy by requiring all recommenders to certify their qualifications, so the assessment of expertise must rely on existing evidence, i.e., the existing recommendations. In order to assess the expertise of recommenders, we therefore first need to define a way to classify content on the Wikipedia, which may then form the basis for our evaluation of each recommender's expertise. The Wikipedia contains articles about all areas of human knowledge, so the classification of Wikipedia content must be broad, but at the same time intuitive, or at the least easy to learn and understand. This indicates that the classification scheme must have a relatively small number of clearly distinct classes.

We have identified 5 criteria (cf. Section 3.1) that a classification scheme must meet in order to be useful in the context of the WRS. Moreover, we have identified a number of different ways to classify Wikipedia articles that are based on existing knowledge classification schemes (cf. Section 3.2). We present an empirical evaluation of 4 of these schemes according to our 5 evaluation criteria. This evaluation identified one of these schemes as significantly better than the other criteria. Based on this evaluation we have extended the WRS, so that it maintains separate trust values for recommenders in each of the classes that they have provided feedback in.

The rest of this paper is organized in the following way: We examine the problem of assessing the expertise of recommenders in reputation and recommendation systems in Section 2, where we also illustrate how the assessment may help improve the accuracy of the trust values used to calculate the ratings of articles. The problem of classifying Wikipedia content and thus the area of expertise of recommenders is examined in Section 3, where we also consider different classification schemes for the WRS. In Section 4, we evaluate 4 of most promising of these evaluation schemes and identify one scheme that satisfies all of our evaluation criteria. We present an overview of the Wikipedia Recommender System and outline the implementation of our extension that evaluates the expertise of recommenders in Section 5. Finally, we present our conclusions and directions for future work in Section 6.

2. Motivation

Many reputation systems provide users with a single rating, generally in the form of a numerical value or a number of stars. The interpretation of these ratings is often implicitly given by the range to which they belong, e.g., a rating of 2 on a scale 1–10 is poor but 4 out of 5 stars is good. These reputation ratings are based on the feedback from other users of the reputation system. This feedback often includes several attributes, e.g., the detailed seller information on eBay consists of the following four attributes "Item as described", "Communication", "Shipping time" and "Shipping and handling charges". We examine some of the problems that may arise when aggregating multiple quality attributes into a single rating in the following.

2.1 Trust Metrics in Reputation Ratings

Users providing feedback may not all be equally competent to evaluate the feed-

back attributes or they may simply have different expectations, e.g., a buyer on eBay may expect overnight delivery as fast "Shipping time" even when ordering items from an overseas seller. It is therefore necessary to evaluate the experience of the user providing the feedback in order to properly calibrate the inclusion of the feedback in the calculation of the aggregated reputation value.

In the context of the Wikipedia Recommender System, an evaluation of the experience of a recommender aims to determine whether she is knowledgeable in the domain of the article and whether she is able to recognise an accurate, complete and concise article. The ratings provided by the first version of the WRS [11),12),14)] capture the second set of qualifications, but the system does not consider the domain of the article for which feedback is given. This means that the ratings from WRS users who have provided good feedback in one domain will carry more weight in all other domains, e.g., a WRS user who has provided good feedback about drag racing is automatically believed when she provides feedback about painters from the Italian Renaissance; this is not necessarily a good idea.

We have identified two ways to establish the expertise of recommenders, either through certification or through an evaluation of the recommender's past performance within each individual domain, but in either case we need to know the domain of all rated Wikipedia articles so that we may determine whether the recommender's expertise applies to that domain. Establishing expertise based on certification requires all recommenders to document their qualifications, e.g., by making certified copies of their diplomas available on their Wikipedia user page. However, this violates both the Wikipedia's policy of allowing anonymous modifications and the privacy of recommenders. Moreover, it introduces the problem of interpreting the value of the different types of qualifications, such as establishing a universal ranking of all the different accredited and non-accredited universities. Finally, it does not allow the incorporation of recommendations from autodidacts. We therefore believe that it is better to base the ratings on an evaluation of the expertise demonstrated by the author's past performance. We propose to do this by classifying rated Wikipedia articles and apply the existing reputation system to the articles within each class.

2.2 Expertise of Recommenders

In order to demonstrate the relevance of evaluating the expertise of recommenders, we need to demonstrate that recommenders have different areas of expertise. In particular, we are interested in determining whether recommenders have particular areas of expertise or whether they are all omniscient. We do this by examining the registration of authors on the Citizendium Work Groups [⋆1]. Authors on the Citizendium have to register explicitly for a work group before they can edit articles belonging to that group. We assume that the authors only wish to register in work groups within their areas of expertise, so that the composition of areas of expertise among Citizendium authors correspond to a self assessment of areas from their areas of expertise. Moreover, we assume that the distribution of areas of expertise in this self assessment also applies among the broader Wikipedia user base. i.e., if Citizendium authors generally register in a few categories, we consider this to be evidence that Internet users are not omniscient and that categorization of their recommendations sense. Our analysis of the Citizendium author registrations is shown in **Fig. 1**.

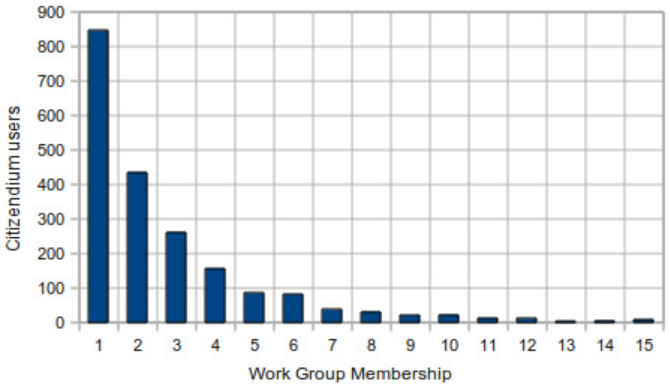


Fig. 1 Registration of Citizendium authors in workgroups.

The figure shows the number of authors who have registered in respectively 1, 2, ... 15 different work groups. There is a clear tendency that fewer authors register for a larger number of work groups, i.e., from 847 authors registered in a

⋆1 The Citizendium is similar to the Wikipedia, but there are only around 2000 registered authors, so it is more tractable than the Wikipedia with more than 11 million registered users.

single work group to 9 authors who have registered in all 15 work groups. This supports our hypothesis that authors of web-based encyclopedia only consider themselves competent in a small number of subject areas.

2.3 Expertise Impact on Trust Metrics

In order to illustrate the benefits of including an evaluation of the recommender expertise in the trust metrics, we consider a scenario where Alice is calculating the weight of a recommendation from Bob about the article "Quantum mechanics". Alice has previously seen 10 recommendations from Bob and she uses her own agreement or disagreement with these recommendations to calculate the weight used in the rating calculation – in trust terminology, Alice is calculating a trust value for Bob based on her prior experiences (good or bad) with Bob. Bob is a typical geek, so he generally provides good feedback in the area of science and technology, but has difficulty assessing articles in religion, social science, arts and the humanities. Alice has seen the following recommendations from Bob (her experience is indicated in a parenthesis following the article name): "Albert Einstein" (good), "Schrödinger's cat" (good), "Ludwig van Beethoven" (bad), "Moon landing" (good), "Tesla Motors" (good), "Rotavirus" (bad), "Karl Marx" (bad), "William Shakespeare" (bad), "Basketball" (good) and "Chicago Bulls" (good). The overall result of the interaction experiences is 6 good interactions and 4 bad interactions, but it is clear that Bob provides good recommendations on the topics of science, technology and sports, while his feedback is less valuable on the topics of psychology, medicine, economics and the arts. The trust metrics implemented in most reputation systems, such as the one implemented in the first WRS prototype [9], simply incorporate the number of good and bad experiences without considering the expertise of the recommenders. **Figure 2** shows the evolution of the trust value as a function of all the experiences [⋆1].

At the time when Alice calculates the rating for the "Quantum mechanics" article, the trust value for Bob is 0.7, which reflects the small overweight in positive experiences when all experiences are considered together.

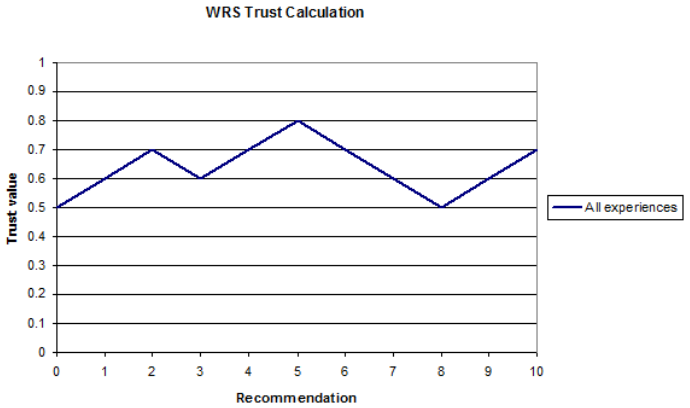


Fig. 2 Trust evolution in the first WRS prototype.

The potential impact of including an evaluation of the expertise of recommenders is illustrated in **Fig. 3** and **Fig. 4** – Fig. 3 shows the impact of a simple classification scheme with only two classes: Science & Technology and Humanities, while Fig. 4 shows the impact of a classification scheme with 6 classes [⋆2]. The two classes shown in Fig. 3 should be interpreted as a class consisting of science and technology and a class that consists of everything that is not clearly natural science; this means that articles on subjects in social science, medicine, arts and religion are classified as "Humanities" in this scheme.

In our scenario, Bob is a stereotypical geek who provides good feedback on anything related to science. In the trust evaluation scheme with 2 subject categories, the trust value for Bob is 0.9 which reflects Alice's 4 previous good experiences with recommendations from Bob in this category. This is the trust value that will be used to decide the weight of Bobs recommendation in the rating for the article on "Quantum mechanics". Bob is also interested in basketball, which means that the effect of his 4 bad recommendations regarding music, medicine, economics and literature is mitigated by his two good recommendations about basketball.

In the trust evaluation scheme with 6 categories shown in Fig. 4, Bob's good recommendations are divided among three classes ("Natural Sciences", "Applied

⋆1 To improve the clarity of the illustration, the figure shows a simplified trust function with trust values in the interval $[0, 1]$, where 0.5 represents the initial values (*unknown*) and the trust value is incremented with 0.1 for each *good* experience and decremented with 0.1 for each *bad* experience.

⋆2 The 6 classes correspond to the top level workgroups in the Citizendium.

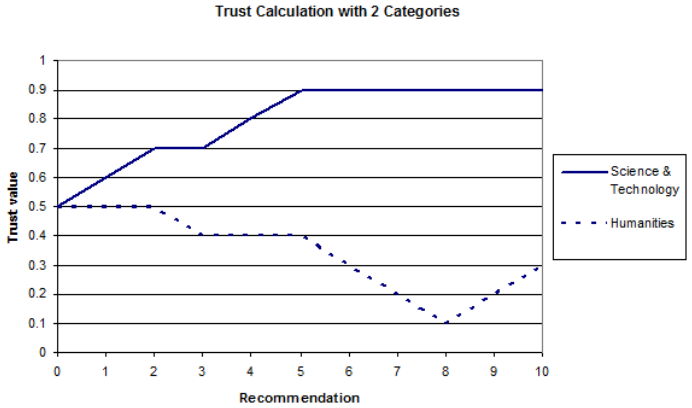


Fig. 3 Recommender expertise evaluation with 2 categories.

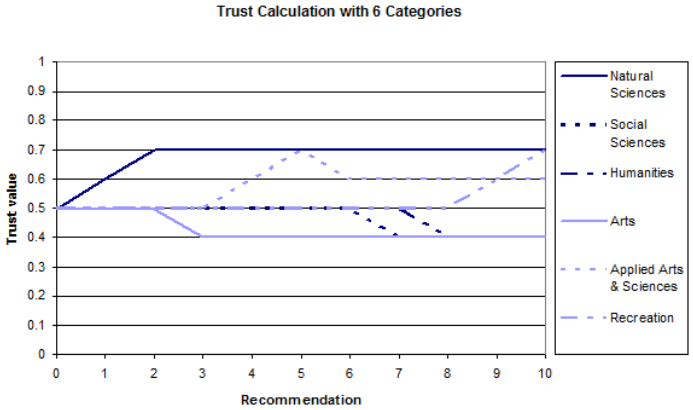


Fig. 4 Recommender expertise evaluation with 6 categories.

Arts & Sciences" and "Recreation") with two good recommendations in each category. This means that the trust value for Bob used in the calculation of the rating for the "Quantum mechanics" article is 0.7, which is no better than the scheme that does not consider the experience of recommenders and significantly lower than the trust value in the scheme with only two classes.

The three different ways to calculate the proposed scenario demonstrate that an evaluation of recommender expertise may significantly improve the precision of the trust value calculation (from 0.7 to 0.9). It also demonstrates the importance of the classification scheme in the evaluation of recommender expertise.

3. Classification of Wikipedia Articles

There are hundreds of classification schemes which are used to classify information. Classification specialists and generally agree that there is no "best" classification scheme, but some schemes are very popular. The factor which determines the popularity of a scheme is the universal adaptation which is relevant to the coverage of different topics. Logical structure and ease of search, i.e., usability and usefulness, are the keys to success.

The fact that there is no "best" classification scheme suggests that none of the existing classification schemes address all the needs for information classification. We conjecture that constructing a single consistent and complete scheme for classification of knowledge will be impossible, because knowledge will often be classified in different ways according to the context, e.g., an article on "Albert Einstein" may be classified as *natural science* by a physics student examining the relativity theory, but classified as *biography* by a history student writing an essay on the great scientists of the 20$^{\text{th}}$ century.

As we propose to use feedback from WRS users to classify the articles, we need to decide on a classification scheme that is easy to use and results in a minimum of ambiguity about what class an article belongs to.

3.1 Classification Scheme Evaluation Criteria

In order to select the most appropriate classification scheme for WRS, we need to define some criteria for evaluation of classification schemes [10]. These criteria should identify the features that are expected from the scheme, but they may also provide valuable input to the development of the trust metrics. We have identified the following criteria for a collaborative classification scheme for Wikipedia articles. A classification scheme should be:

Intuitive People should find it easy to classify an article.

Complete There should be a class for every article.

Concise There should be as many articles in every class as possible, so there should only be a relatively small number of distinct classes.

Unambiguous People should generally agree on the classification of articles.

Useful The classification should improve the ratings of WRS.

The first criteria should be relatively self-explanatory; if users of the WRS find it difficult to classify articles they are unlikely to go through the bother of providing feedback, which means that there may not be sufficient recommendations to calculate an accurate rating (this may cause other WRS users to loose interest and start a vicious circle). The completeness criteria means that there should be no articles that are impossible to classify. This means that there must be enough classes to encompass the entire body of human knowledge. The conciseness criteria limits the number of classes in the scheme, which increases the likelihood that a recommender of an article has already rated other articles in that class, i.e., this helps reduce the *cold start problem* [20]. A concise classification scheme also makes it easier for people to remember the classes, i.e., it also improves *intuitiveness*. As mentioned above, the classification of an article may depend on the context of the classifier, so the classification scheme should facilitate consistent classification of articles regardless of the classifier's context, e.g., the scheme should not include separate classes that are very similar, e.g., *rocket science*, *space flight* and *interplanetary travel*. We do not, however, believe that a collaborative classification scheme for Wikipedia articles will provide a complete and consistent classification of all Wikipedia articles. We therefore propose to evaluate the unambiguity criteria empirically through the experiment described in Section 4. Finally, the classification of articles should provide better ratings to users of the WRS, i.e., the classification scheme should provide *useful* results.

3.2 Classification Schemes

We have identified two different ways to classify Wikipedia articles. We may either define an *internal classification scheme* that tries to infer a classification from the information that is already available in the Wikipedia, such as Portals or Wikipedia Categories, or we may define an *external classification scheme* that relies on feedback from WRS users to categorise the article – this means that recommendation ratings must be interpreted in the context that the recommender specifies.

3.2.1 Internal Classification Schemes

The Wikipedia includes several disparate schemes to classify articles. In the following, we focus on the two main efforts based on *portals* and *categories*.

Table 1 Wikipedia's contents/portals.

General reference	History and events	Philosophy and thinking
Culture and the arts	Mathematics and logic	Religion and belief systems
Geography and places	Natural and physical sciences	Society and social sciences
Health and fitness	People and self	Technology and applied sciences

3.2.1.1 Portals

A portal, or Wikiportal, on the Wikipedia serves as an entry-point to Wikipedia content within a topic area. Portals vary from very broad coverage, such as the *History* portal, down to very specific topics, such as the *Led Zeppelin* portal. The Wikipedia has 12 main portals as shown in **Table 1**, but there are currently a total of 583 portals on the Wikipedia [⋆1]. The Wikipedia portals are hierarchically structured, so it is possible to enter a portal and find a selection of articles and sub-portals, but it is generally impossible to enter an article and find out which portal it belongs to. Moreover, articles may be reachable from several portals, which introduces problems if we wish to assign a unique category to each article. Finally, a portal is not a complete enumeration of articles belonging to the category, or topic, of the portal. This severely limits the use of portals as a means to determine the category of an article, because we cannot be sure to find an article even if we traverse all links on pages and sub pages reachable from the main portals. Using the portals to determine the category of an article is therefore going to be computationally difficult, perhaps even impossible, and from an overall perspective will yield incomplete results.

3.2.1.2 Wikipedia Categories

Categories is another hierarchical scheme, that have been introduced to allow authors to classify articles in the Wikipedia. The Wikipedia rules state that each article should belong to at least one category, so it is fair to assume that all articles have at least one category, but most articles will have more than one category, e.g., the "London" article belongs to 8 categories. This means that categories cannot be used directly to classify articles, but each article has a set

⋆1 Information about the Wikipedia portals can be found on the Wikipedia: http://en.wikipedia.org/wiki/Portal:Contents/Portals, visited 7 April, 2010.

of categories and it is possible to traverse bi-directional links in the article all the way to the root category (and back down again). It should be possible to follow all the category links to their root category and use these to define a classification of the article. There is, however, one major problem with this solution, which is that categories are socially annotated, i.e., they are all created and maintained by Wikipedia users. While the initial idea behind the categories were that they should be shaped into a tree-like structure, the actual structure has mutated into a more general graph structure, e.g., each leaf may have several parents and there may even be cycles, so some categories are their own grandparents. Finally, the set of categories is not fixed, so new categories will be created as the Wikipedia expands. These new categories may have a significant overlap with existing categories, which means that an evaluation of recommender expertise would have to transfer (some of) the expertise demonstrated in the context of older categories to each of the new, overlapping categories. It is not clear to us how this may be achieved in practice and, when added to the other difficulties outlined above, it is difficult to see how categories may be used to provide a simple unambiguous classification of the articles in the Wikipedia.

3.2.2 External Classification Schemes

As mentioned above, social annotations are often dynamic, which makes them unsuitable for the definition of a classification scheme for a dynamically growing set of articles. However, once the classification scheme has been defined, social annotations may be used to assign categories to articles. We examine the problem of defining a simple and intuitive classification scheme for the Wikipedia, which is intended to cover all areas of human knowledge.

3.2.2.1 Wikipedia Portals and Categories

We briefly revisit the idea of using the existing set of Portals or Root Categories to define the classification scheme. As mentioned above, both of these schemes are dynamically growing, i.e., new Portals and new Root Categories maybe introduced into the system. This means that neither provide the stable reference structure that we need to support our evaluation of recommender expertise. Similar problems exist in the Citizendium [4),⋆1], which defines 6 "general areas" and a number of work groups within each of these areas. The general areas in the Citizendium are shown in **Table 2**.

Table 2 General areas of information on the Citizendium.

Natural Sciences	Social Sciences	Humanities
Arts	Applied Arts and Sciences	Recreation

We have therefore decided to focus on existing classification schemes used in libraries, which have been designed to classify all areas of human knowledge. In particular, we examine: the "Library of Congress Classification", the "Universal Decimal Classification" and the "Dewey Decimal Classification". It is important to remember that a classification system used in this context has different requirements than it has in a library. We need a system that covers the entire spectrum of knowledge, but in a simple and unambiguous way, i.e., the system should not require a librarian education to operate it.

3.2.2.2 Library of Congress Classification

The Library of Congress Classification [5)] (LCC) is developed by one specific library, the US Library of Congress. It is in widespread use among research and academic libraries and as such qualifies for consideration. The system contains 21 classes and new classes have been added as needed, which has led to much criticism because of a lack of a sound theoretical basis, e.g., some unusual sciences have their own categories, such as Military and Naval sciences. Another problem is that it is regionally biased towards the US, which can be seen by the fact that there are separate categories for *world history* and the *history of the Americas*. The size and peculiarity of the classification scheme means that the LCC is not considered sufficiently intuitive for the WRS.

3.2.2.3 Universal and Dewey Decimal Classification

The Universal Decimal Classification [23)] (UDC) is derived from the Dewey Decimal Classification [15)] (DDC), so we discuss both here. The UDC uses a complex system of additional symbols to indicate special aspects or relationships of a subject. Both systems have 10 well defined base classes [⋆2] which makes them

⋆1 The Citizendium is a collaboratively edited encyclopedia on the web, with a more strict editorial policy than the Wikipedia. It was started by Larry Sanger, the co-founder of Wikipedia, as a Wikipedia fork which aims to address the problem of reliability and quality by only allowing users with real-name registration to edit Citizendium pages [19)].

⋆2 The UDC only 9 of the 10 classes leaving category "4" vacant.

Table 3 The Dewey Decimal Classification.

Class	Description
000	Computer science, information, and general works
100	Philosophy and psychology
200	Religion
300	Social sciences
400	Languages
500	Science and Mathematics
600	Technology and applied science
700	Arts and recreation
800	Literature
900	History and geography and biography

interesting from a usability perspective and neither have the regional bias of the LCC. While either of these systems satisfy our requirement for a classification scheme, we believe that the DDC is more descriptive, so we have decided to use this scheme in the WRS. The DDC class table (**Table 3**) is as follows: Using a complete classification scheme, such as the DDC, means that WRS users may consult the reference definition in case of uncertainty.

3.2.2.4 Open Directory Project – Dmoz

The Open Directory Project [16], commonly known as Dmoz (from directory.mozilla.org, which was its original domain name), intends to provide a comprehensive directory of the web. The continual growth of the Internet makes it increasingly difficult for search engines to return relevant information on the first page and commercial directory sites cannot keep up with submissions, so the quality and comprehensiveness of these directories have also suffered.

The Open Directory project is owned by Netscape, but the directory is maintained by a community of volunteer editors who are responsible for the inclusion of sites into the web directory. Everybody can submit a site to the Open Directory, which is then reviewed by a category editor before it is listed in the directory. Dmoz also encourages everyone who genuinely wishes to participate in the building of the Open Directory to volunteer as a category editor. Dmoz is used by well known search engines and portals, including Google, AOL Search, Netscape Search, Lycos, DirectHit and others.

Dmoz defines a single directory of Internet content based on a hierarchical ontology scheme for organizing site listings. Listings on a similar topic are grouped into categories, which can then include smaller categories. Dmoz defines 15 top level categories which are listed in **Table 4**.

Table 4 Open Directory Project top-level categories.

Arts	Business	Computers
Games	Health	Home
Kids and Teens	News	Recreation
Reference	Regional	Science
Shopping	Society	Sport

4. Classification Scheme Evaluation

In our evaluation of classification schemes, we only consider external classification schemes and we focus our evaluation on four schemes that we believe provide a representative selection of classifications with a varying number of categories. These schemes are: the Wikiportals scheme, the Citizendium scheme, the Dewey Decimal Classification scheme and the Open Directory Project scheme. All of these schemes have been designed to classify the full body of human knowledge, so they all satisfy the completeness criteria. Despite a variation in the number of categories, none of the schemes have more than 15 categories, so they also satisfy the conciseness criterion. This means that our evaluation will primarily address the unambiguity and intuitiveness criteria.

In order to empirically identify the classification scheme that best satisfies these requirements, we conducted an online survey, which was open for 5 weeks in the summer of 2010. During that period, 130 people from different countries and continents participated in the survey (this eliminates most cultural bias). Participants were asked to read 4 Wikipedia articles and categorize them according to one of the classification schemes. The classification scheme was selected by the system based on the time of day, but the schedule for presenting the different surveys was changed every day, so that for any given time of day, participants in the same time-zone had an equal chance of encountering each of the classification scheme, i.e., participants who returned back from work at 7 pm would have an equal chance of being asked to use each of the classification schemes – this would

only depend on the day in the week that they decided to answer the survey. The first part of the survey is primarily designed to determine which of the classification schemes best meet the unambiguity criteria. We therefore selected articles with a high potential for ambiguity; all the articles are directly accessible from more than one of the top level portals in the Wikipedia.

When article categorization was completed, people were asked to evaluate their experience with answering the survey. These questions are primarily designed to assess how intuitive the participants found the classification scheme.

4.1 Wikiportals

There are 12 top-level portals in the Wikipedia (these are often called the Wikiportals) which define entry points into different areas of knowledge in the Wikipedia. These entry points may be interpreted as leading into separate branches of the Wikipedia and may therefore form the basis of a classification scheme. The categories in this scheme are shown in Table 1. The results from the survey are shown in **Fig. 5**.

The figure shows the percentage of survey participants, who classified the article in that particular category, e.g., more than 40% classified the article on "Albert Einstein" as *People and self*. The red line indicates the base rate that must be assumed if the classification is based on guessing, i.e., if there are no obvious choices for an article in the classification scheme.

The figure shows that there is only a clear classification for Article 3, where 66% classified "Rotavirus" as *Health and fitness* and only one other category (*Natural and physical sciences*) is above the base rate (with 20%). The article on "Albert Einstein" has one class (*People and self*), which scores significantly higher than any other class, but there are three classes that score sufficiently much higher than the base rate to merit consideration. In both the articles on "Moon landing" and "Basketball" two classes score significantly higher than any other class, but in both cases the difference between these two classes is too small to allow a clear classification. We therefore conclude that the Wikiportals scheme does not completely satisfy the unambiguity criteria.

Figure 6 shows answers to the questions designed to reveal how intuitive the scheme is.

The figures show that a clear majority of participants agreed that it was easy

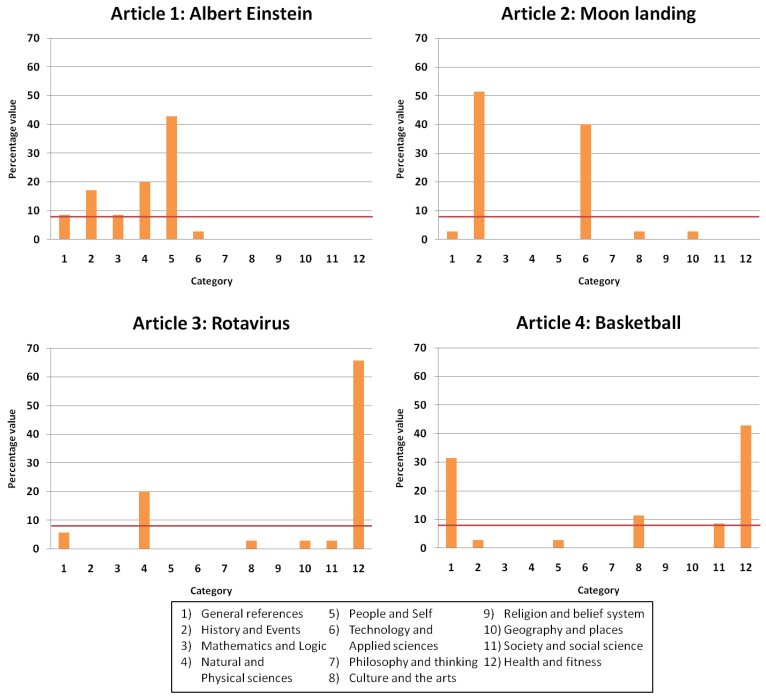


Fig. 5 Articles classification results with Wikiportals.

to find the right category and that the names of the categories make sense (Question 1 and Question 4). The answers to the question about whether it is easy to understand the logic behind the classification is less decided (Question 2) and there is a significant number of participants who moderately disagreed that the categories are specific enough (Question 3). These answers indicate that the participants are generally satisfied with Wikiportals classification scheme and that they found it meaningful and easy to use. The moderate disagreement in Question 3 is also consistent with the answers in the article classification survey, because there was only a clear classification of the "Rotavirus" article, so despite the general satisfaction with the scheme some users found that the categories were not specific enough. We therefore conclude that the scheme appears intu-

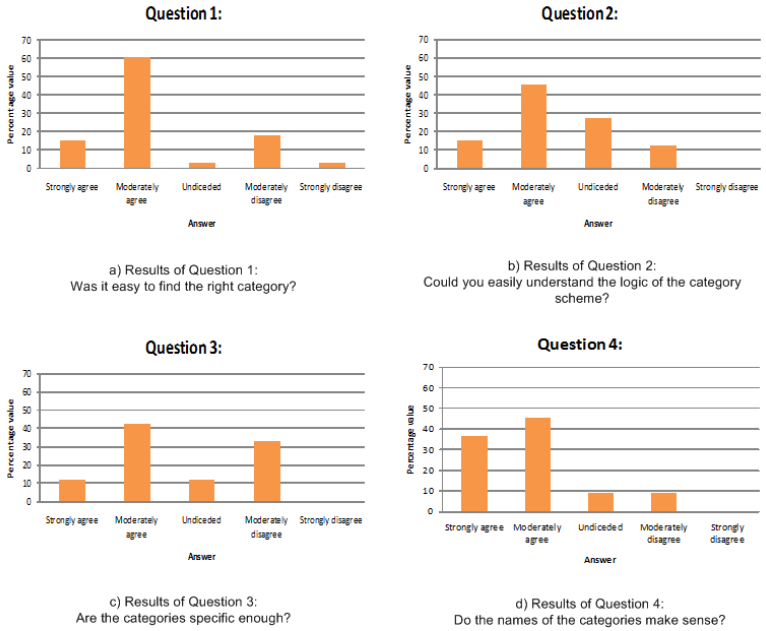


Fig. 6 Answers into questions about Wikiportals.

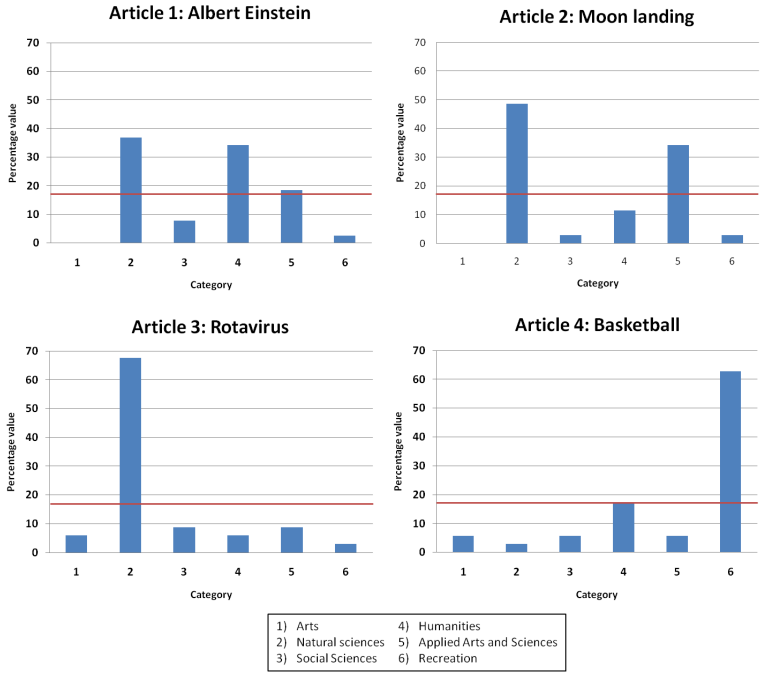


Fig. 7 Articles classification results with Citizendium.

itive to the participants and that most people are satisfied with the categories, but the experiments indicate that users have some difficulties using the classes in practice.

4.2 Citizendium

The Citizendium classification scheme contains only 6 top-level categories, which is half the number of classes in the Wikiportals scheme. The result of our evaluation of the Wikiportals scheme indicated that the 12 classes were not specific enough to classify the four articles, so we expect that the results from the Citizendium scheme would show a significantly higher degree of ambiguity and that fewer participants are satisfied with the scheme. The results of the article classification with the Citizendium scheme are shown in **Fig. 7**.

The figures show that the smaller number of classes actually made it easier to agree on a class and there was a clear classification for two articles (both articles on "Rotavirus" and "Basketball" got a single class with more than 60% of the answers and no other class got significantly more than the base rate). The results revealed that the majority of participants assigned the top-level category *Natural Sciences* to the article on "Rotavirus" (Fig. 7, Article 3) and *Recreation* to the article on "Basketball" (Fig. 7, Article 4). The classification of the two other articles was more evenly distributed over the categories, but in both cases two categories were selected by more participants than the remaining categories. These results, however, are not as marked as the results from the Wikiportals scheme and all four articles have been classified in 5 of the 6 categories by at least one participant. We therefore conclude that, despite getting better results than the Wikiportals scheme, the Citizendium scheme does not completely satisfy the unambiguity criteria.

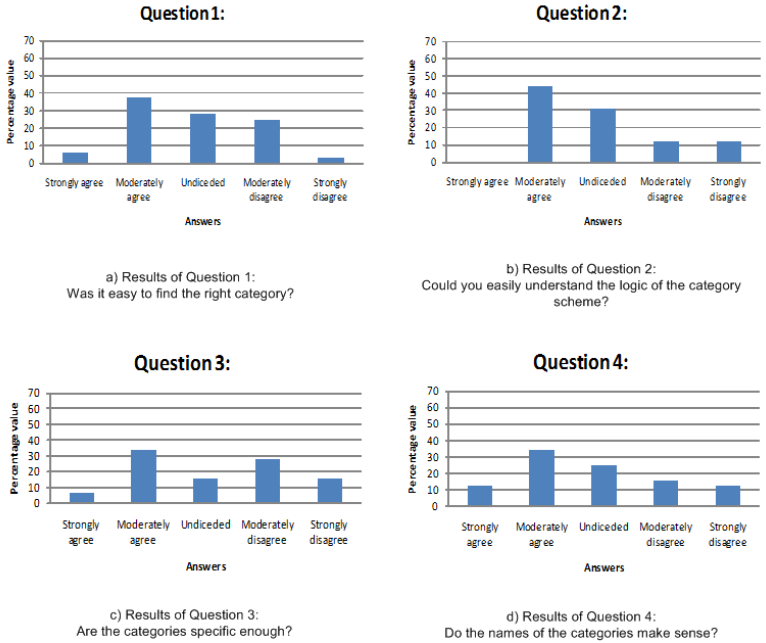


Fig. 8 Answers into questions about Citizendium.

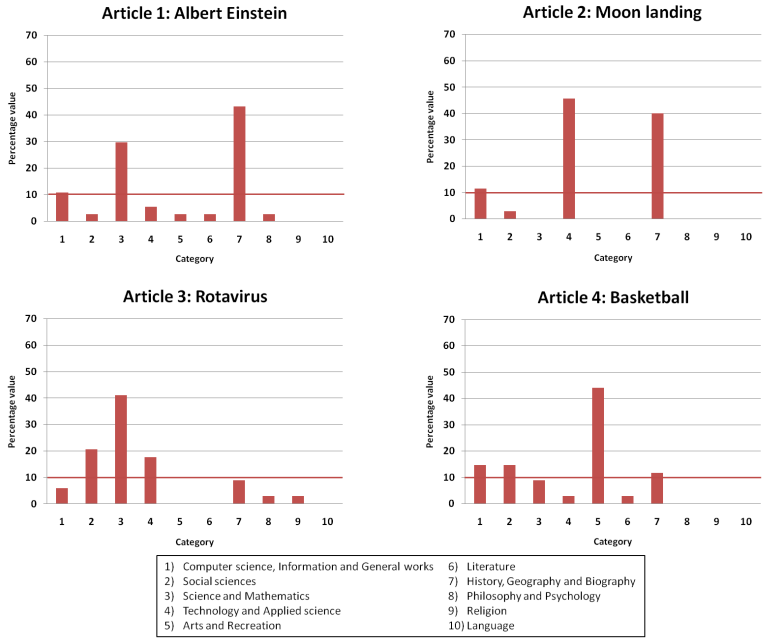


Fig. 9 Articles classification results with Dewey Decimal Classification.

Our evaluation of how intuitive the participants found the scheme to be is shown in **Fig. 8**.

The figures show that most of the participants found it more difficult to find the right category and understand the logic behind the classification scheme than was the case with the Wikiportals scheme. Moreover, we find that 16% disagreed strongly with the question "are the categories specific enough" (Fig. 8 c), so they obviously had difficulty finding an appropriate class for one or more of the articles. Finally, fewer of the participants found that the names of the categories made sense and more than 10% found that they did not make sense at all.

The conclusion of this survey is that the participants were generally not satisfied by the smaller number of categories and found it more difficult to use, so the scheme does not satisfy the intuitiveness criteria.

4.3 Dewey Decimal Classification

The Dewey Decimal Classification (DDC) scheme is one of the most widely used classification schemes in the World. The scheme is comprised of 10 top-level categories with an enumeration from 0 to 9. In his Masters Thesis [13], Thomas Lefévre selected the DDC scheme as the most suitable scheme for the classification of the entire body of human knowledge and implemented it as the main classification scheme in a version of the WRS. The scheme, however, was never properly evaluated, so we decided to subject it to the same evaluation as the other schemes examined in this paper; the results are shown in **Fig. 9**.

The figures show that none of the Wikipedia articles were clearly classified. In all cases the article was mainly classified in 2 or three categories, but only the article on "Basketball" had a single class (*arts and recreation*) that received significantly more replies than the others.

To explain this distribution of categories, it is worth remembering that articles about "Albert Einstein" and the "Moon landing" could belong to several categories, such as *Computer science, information and general works* and *Sciences and Mathematics* that have overlapping definitions which could explain the poor result. Moreover, examining the results of how people classified articles on "Rotavirus" and "Basketball" we are surprised that 15% of the participants selected the category *Computer science, information and general works* for the article on "Basketball" and that 15% of the participants classified it as *Social Science* while only 44% selected the category *Arts and Recreation* – which we consider the correct classification. The high number of classifications in the category *Computer science, information and general works* may be explained by the "miscellaneous/catch-it-all" nature of this category. Such "miscellaneous" classes may help guarantee completeness, but our evaluation indicates that they are dangerous to include in a collaborative classification scheme and should be avoided. The reason for the low number of correct classifications of the article on "Basketball" is probably that the first word *Arts* misled participants and made them look for another category. The overall shape of the histograms produced by the DDC scheme and the Wikiportal scheme are quite similar. Both schemes have one article that was most easily classified in a single category (the Wikiportal scheme clearly classifies the "Rotavirus" article as *Health and fitness* while the DDC scheme has a less clear classification of the "Basketball" article as *Arts and Recreation*), two articles that are predominantly classified in two categories (in the Wikiportals scheme the article on the "Moon landing" is classified as either *History and events* or *Technology and applied sciences* and the article on "Basketball" is classified as either *General references* or *Health and fitness* while in the DDC scheme the article on the "Moon landing" is either as *Technology and applied science* or *History, geography and biography* and the article on "Albert Einstein" is either classified as *Sciences and Mathematics* or *History, geography and biography*) and one article that has a relatively high number of classifications in several schemes (in the Wikiportals scheme the article on "Albert Einstein" has a significant number of classifications in 5 categories while in the DDC scheme the article on "Rotavirus" has a significant number of classifications in 4 categories). Overall, the results of our evaluation of the Dewey Decimal Classification

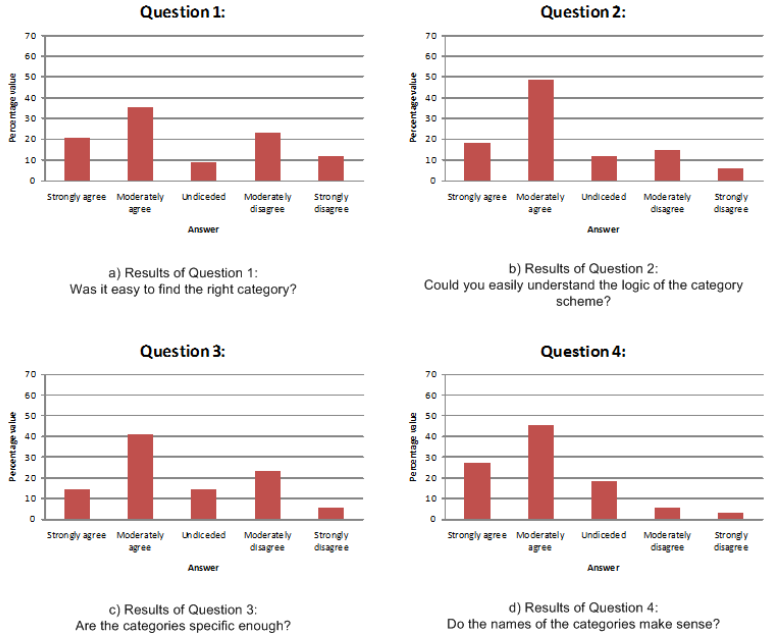


Fig. 10 Answers into questions about Dewey Decimal Classification.

scheme has been disappointing and we have to conclude that the scheme dos not satisfy the unambiguity criteria.

The results of our evaluation of how intuitive the participants found the Dewey Decimal Classification scheme are shown in **Fig. 10**.

The figures show that the participant's satisfaction with the scheme lies somewhere between the Wikiportals scheme and the Citizendium scheme. This appears natural, because the scheme has 10 categories which is in between the 12 classes in the Wikiportals scheme and the 6 classes in the Citizendium scheme.

Overall, the evaluation of this scheme lies somewhere between the Citizendium scheme and the Wikiportal scheme with respect to both unambiguity and intuitiveness. This suggest that the DDC is a good compromise candidate.

4.4 Open Directory Project

The goal of the Open Directory Project (Dmoz) is to provide a collabora-

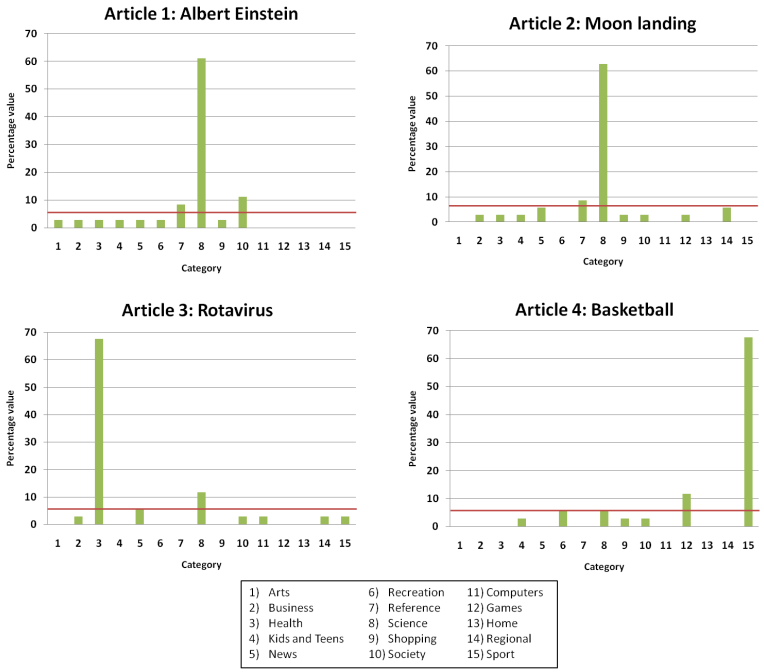


Fig. 11 Articles classification results with Open Directory Project.

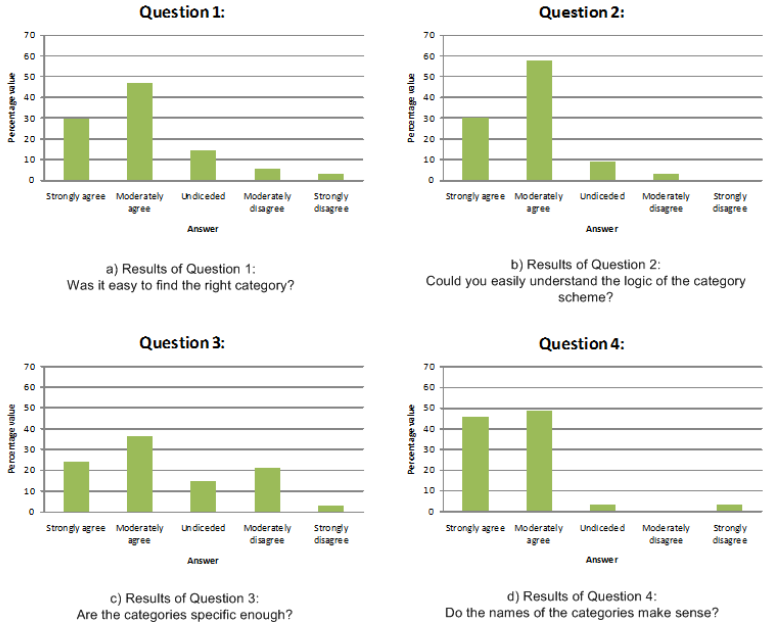


Fig. 12 Answers into questions about Open Directory Project.

tive classification of all pages on the World Wide Web. Dmoz is a hierarchical classification scheme, where web-pages are classified into one of 15 top-level categories, which are divided into sub-categories, which may be divided into sub-sub-categories etc. For the purpose of classifying articles in the Wikipedia we only consider the 15 top-level categories. The result of our classification survey are shown in **Fig. 11**.

The figures show that the Dmoz classification scheme produced remarkably unambiguous results on the four articles selected for our survey. All articles have been classified in a single category by at least 60% of the participants (this is around 10 times the base rate) and there are generally only one other category that receives more than the base rate, but never more than double the base rate.

The articles on "Albert Einstein" and the "Moon landing" which are generally classified in several categories were classified reliably within the Dmoz scheme (61% of the participants classified the "Albert Einstein" article as *Science* and 63% of the participants agreed that the "Moon landing" article also belongs to *Science*). The articles on "Rotavirus" and "Basketball" are classified even more reliably. "Rotavirus" scores 68% in the category *Health* and "Basketball" scores 68% in the category *Sport*. This unambiguity can be explained by the higher number of categories, which means that the scheme is specific enough to classify articles and do not have overlapping categories. We therefor conclude that the Dmoz scheme completely satisfies the unambiguity criteria for the surveyed articles.

The results of our evaluation of how intuitive the participants found the Open Directory Project scheme are shown in **Fig. 12**.

The figures show that the participants generally agree that the scheme is easy

to use and that it is easy to understand the logic behind the Dmoz scheme. In all figures, the degrees of moderate and strong agreement are significantly higher than in any of the other classification schemes (only the Wikiportal scheme have similar but poorer results). Moreover, there was a significant majority of participants who agreed that the category names make sense (Fig. 12 d). This is one of the key questions regarding intuitiveness, reflecting the ability to choose the proper category easily and correctly. We therefore conclude that the Dmoz scheme satisfies the intuitiveness criteria.

4.5 Discussion

Our evaluation of the first three classification schemes suggested a proportionality between the number of categories in the scheme and both the ambiguity and intuitiveness of the classification. The Citizendium has the smallest number of categories, but the highest number of unambiguously classified articles and the lowest degree of satisfaction about usability, while this is the opposite for the Wikiportals scheme. Our expectations would therefore be that the Dmoz scheme, with 15 categories, would be less intuitive and less unambiguous than all the other schemes, but the experiments showed that the Dmoz scheme was completely unambiguous and scored highest on the user satisfaction questions in the survey.

There are a few issues that must be kept in mind when interpreting the results of our survey. First of all, the participants in the survey were selected among friends and acquaintances, mailing list members, online gaming guilds and Facebook groups. Although the number of participants in the survey (130) is sufficient to provide significant results, the population is likely to over represent young students or academics with a higher than average knowledge about computers. The participants may therefore find it easier to understand and utilize a higher number of categories in the classification scheme. We do, however, not believe that this has an impact on our evaluation results, because participants were generally more satisfied with the schemes that had more categories (there may be a limit to the number of classes that can be used, but we believe that it will more likely be a problem of screen resolution rather than mental capacity that imposes this limit). Secondly, the number of articles surveyed in our evaluation is relatively small (4 out of more than 3.4 million articles in the English language version of the Wikipedia), but the findings of this experiment are consistent with a preliminary evaluation of classification using the DDC scheme [8),13)]. We do not expect to be able to cover a convincing fraction of the Wikipedia in any controlled experiment, so we believe that the real test of the Dmoz classification scheme will be through practical use.

The evaluation presented in this paper does not address the usefulness criteria. We are in the process of finalising a distribution of the WRS that can be released to the public which will hopefully provide us (and anyone else who is interested) with a large repository of empirical data. When enough people have adopted the WRS, we will be able to run trace based simulations of the WRS trust evolution function, which will allow us to determine the usefulness of the Dmoz classification scheme in the WRS (compared to not having a classification scheme).

4.6 Summary of Evaluation

In this section we have analyzed 4 different classification schemes in order to identify the best classification scheme for the Wikipedia Recommender System (the criteria for evaluating the classification scheme were defined in Section 3.1).

Our evaluation is based on an online survey that was answered by 130 participants from around the world. The result of this survey showed that the Open Directory Project (Dmoz) classification scheme satisfies both the unambiguity and intuitiveness criteria for the surveyed articles. We have therefore decided to implement this scheme in the current prototype of the WRS.

5. The Wikipedia Recommender System

The Wikipedia Recommender System (WRS) has been designed to integrate with the existing Wikipedia without requiring modifications to the MediaWiki installation or the underlying Wiki engine. The design is based on a generic architecture for the integration of a reputation system in a large web-based legacy system, such as the Wikipedia [12)]. Before describing how this architecture is implemented in the WRS, we provide a brief overview of how the WRS works in practice.

5.1 WRS Overview

The WRS is mostly implemented in a web-proxy, which mediates all commu-

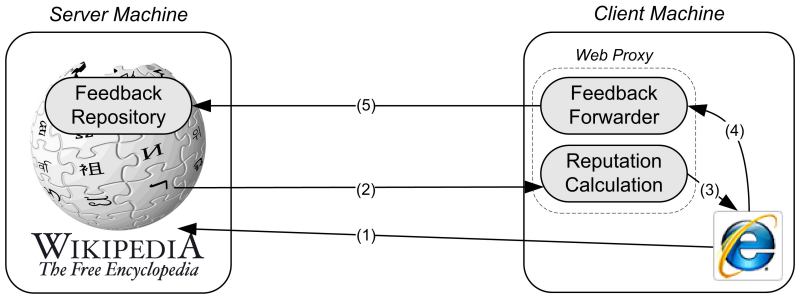


Fig. 13 Overview of the Wikipedia recommender system.

nication between the user's browser and the Wikipedia. Recommendations are stored in the Wikipedia itself, but the recovery and distribution of recommendations, calculation of reputation ratings and formation and evolution of the user's trust in recommenders are managed by the web-proxy. This is illustrated in **Fig. 13**, where the proxy executes on the user's own computer along with the browser. The browser must be configured to use the local web proxy (this is how users opt in), which intercepts requests to the Wikipedia (1). The proxy retrieves the article from the Wikipedia (2) along with the feedback, which is used to calculate the reputation score for the article. The page[⋆1] is rewritten to include the reputation score and forwarded to the browser (3). The user now has an indication of the quality of the article and may decide to provide feedback regarding the quality of the article and the utility of the reputation rating (4). The user's indication of the utility of the score is used by the proxy to build trust in the recommenders who recommended this article and the user's own rating is stored in the feedback repository in the Wikipedia (5).

5.2 WRS Architecture

The WRS is based on our general architecture for integrating reputation systems with legacy applications, which identifies the following main components: a *Feedback Repository* which stores user feedback, a *Reputation Calculation* component which calculates reputation ratings based on data from the Feedback Repository, an *Identity Management* component which verifies the, possibly virtual, identities of feedback providers[⋆2] and, finally, an *Interception Mechanism* which mediates communication between clients and servers and makes reputation ratings and reputation data available where they are needed.

5.2.1 Feedback Repository

The implementation of the Feedback Repository is based on the observation that everyone can edit the Wikipedia, so we can store user feedback in the Wikipedia itself. We have therefore created a special Wikipedia user and maintain the Feedback Repository on the user page of that user. The feedback consists of a recommendation which includes the recommender's rating of the article, the recommender's choice of category for the article and the recommender's signature.

5.2.2 Identity Management

The Identity Management component is used to verify recommendations, by downloading the recommender's public-key from his Wikipedia user page and validate his signature on the recommendation. When downloading the public-key, the WRS must ensure that the key has been added by the owner of the user page, i.e., that the Wikipedia user name included in the recommendation is equal to the Wikipedia user name who uploaded the public-key. Using the Wikipedia for key distribution, means that we support the same degree of anonymity as the Wikipedia.

5.2.3 Interception Mechanism

The Interception Mechanism in the WRS requires the ability to rewrite the content read from the Wikipedia (to insert recommendations for the user) and to capture and store the feedback from the clients. A simple way to do this is therefore to insert a web-proxy between the user and the Wikipedia.

5.2.4 Reputation Calculation

The proxy also implements the Reputation Calculation module, which calculates the rating for a given article based on all the recommendations for the current version of that article. The rating is calculated as an average of the rat-

⋆1 Articles in the Wikipedia are contained in web pages, so we generally use the term *article* to refer to the logical content and *page* to refer to the physical data structure.

⋆2 Identity Management is only relevant for reputation systems that implement trust metrics, i.e., it is not required for a simple summation based system, such as the one used on eBay.

ings in the recommendations weighted by the user's trust in the recommender. The rating calculator inserts an applet in the Wikipedia page, which displays the overall rating for the article and solicits feedback from the user. When the proxy receives the feedback from the user, it updates the trust values for all the recommenders who rated the article, which are then stored locally, and creates a recommendation for the article which is uploaded to the relevant location in the Wikipedia. The update of trust values calculates both the trust value for the recommender that will be used in the next interaction and the user's dispositional trust (the trust dynamics) for each recommender. Both of these values are calculated as a function of the difference between the number of positive and negative experiences with that recommender (an interaction where the user agrees with the recommender's rating count as a positive experience, but if they disagree it counts as a negative experience) [9),11)]. The proposed evaluation of the expertise of recommenders is intended to refine this notion of positive and negative experience and it to the category of the article.

5.3 Extending the WRS to include Categories

We have implemented a new version of our prototype [18)], which extends the WRS to include an assessment of the expertise of recommenders according to the classification defined in Section 3. As mentioned above, users are now expected to provide both a rating and a category for the article, when they return feedback to the WRS. This means that there are now two separate types of feedback that must be considered by the WRS and the trust metric must, in some way, reflect the recommender's ability to provide reliable feedback of both types. It seems obvious, however, that it is more important that a recommender is able to determine that the article is accurate, complete and well written, so we consider the rating metric the *primary parameter* and the category rating the *secondary parameter* when the WRS updates the trust value.

The introduction of the second type of feedback means that there are now four separate cases that must be considered when the user and recommender ratings are compared, because the may agree or disagree about both ratings and categories, this is illustrated in **Fig. 14**. The two cases where the user and the recommender agree on the category are covered by the trust dynamics implemented in the first WRS prototype, so we only need to define appropriate trust

Rating	Category
Agree	Agree
Agree	Disagree
Disagree	Agree
Disagree	Disagree

Fig. 14 Outcome of interacions.

dynamics for the two other cases. This is an intriguing problem, because both the rating and the category are subjective values. We therefore propose to examine all the other recommendations for the article in order to determine if there is a majority among the other recommenders who support either category (if there is no clear majority, the user carries the deciding vote). We wish to define a decision function that corresponds to human intuitions. We therefore believe that it is reasonable to say that if the majority agrees with the user, this is clear evidence that the recommender is considered to be wrong, but if the majority agrees with the recommender, it must be considered that she could be right. We examine the two cases in greater detail in the following.

5.3.1 Agreement on rating, disagreement on category

Both user and recommender agree on the quality of the article, but at least one of them is wrong about the category, which suggests some problem with the comprehension of the article. However, they both agree on the apparent qualities of the article, which we consider the primary parameter, so the overall interaction is considered positive. In order to reflect the problem with comprehension of the article in the updated trust value, we introduce the notion of semi successful interactions for which a value of $+\frac{1}{2}$ positive interaction seems appropriate.

5.3.2 Disagreement on rating and category

When the user and the recommender disagrees about everything, we need to consider the majority of the other recommenders regarding the category of the article (in short *the majority*). If the majority agrees with the user, the recommender has severely misunderstood the article, so it seems appropriate to penalize him more severly. We therefore consider the recommendation as evidence for a $-\frac{3}{2}$ "positive" interaction. If the majority agrees with the recommender, the recommendation is obviously provided in a different context and should be con-

Rating	Category	Impact
Agree	Agree	1
Agree	Disagree	$\frac{1}{2}$
Disagree	Agree	-1
Disagree	Disagree without majority	$-\frac{1}{2}$
Disagree	Disagree with majority	$-\frac{3}{2}$

Fig. 15 Impact of categorization on trust dynamics.

sidered on its own merrit, i.e., the rating might have been right if the they had agreed on the category. It does not, however, change the fact that the rating is wrong in the user's opinion, so we consider the recommendation as evidence for a $-\frac{1}{2}$ "positive" interaction.

It is important to determine what constitutes a majority. In the mathematical sense it means more than 50%, but that seems inconclusive and unconvincing when deciding on the severity of penalties. We have performed a preliminary survey of Wikipedia user's ability to categorize articles according to our classification scheme [13], which indicates that a majority of more than 80% appears to be safe, even for articles that a small element of ambiguity about the category. Later in this paper (cf. Section 4), we present an evaluation of four classification schemes that indicate that an appropriate classification scheme may ensure that there will be a majority of more than 60% for most articles (the evaluation showed a majority of more than 60% for all the surveyed articles using the best classification scheme – these articles were all selected because of possible ambiguity).

5.3.3 Summary

The evaluation of the expertise of recommenders should have the following impact on the number of "positive" interactions used in the trust evolution function. The first and the third line in the table above correspond to the first prototype of the WRS. The second line displays the effect when the user agrees with the recommender on the rating but disagrees on the category. The two last lines show the impact of the interaction when the user and the recommender disagrees on both rating and category. Line 4 shows the case where the majority agrees with the category of the recommender and line 5 shows the case where the majority agrees with the user (**Fig. 15**).

6. Conclusions

In this paper we examined the problem of assessing the expertise of recommenders in the Wikipedia Recommender System. We illustrated how classifying the articles in the Wikipedia according to a relatively small set of categories and maintaining separate trust values for recommenders in each category might improve the trust values that affect the ratings in the WRS (cf. Section 2.3). There is, however, no complete and consistent classification of Wikipedia articles, but there are a number of potential classification schemes that may be appropriate.

In order to identify the best classification scheme to use in our assessment of recommender expertise, we examined a number of classification schemes that could be considered for the WRS. We identified 4 classification schemes that are commonly used to classify information and evaluated them according to 5 criteria that we presented in Section 3.1, namely that the scheme should be: intuitive, complete, concise, unambiguous and useful. The evaluation consists of an online survey with 130 participants from around the world. We argue that all of the surveyed schemes have been designed for general information classification, so they are all complete. Moreover, none of the schemes have more than 15 classes, so we also consider them to be concise from a usability point of view, but we have not evaluated whether the schemes with a high number of categories are sufficiently concise to avoid significant cold start problems. The evaluation clearly showed that the 15 top-level classes defined in the Open Directory Project (Dmoz), was the only classification scheme to satisfy the unambiguity and intuitiveness criteria. We therefore decided to implement this scheme in the current version of the WRS.

There are two important issues that have not been addressed by our evaluation. First of all, we have not demonstrated that the Dmoz scheme is sufficiently concise to avoid cold start problems. Secondly, we have not evaluated any of the schemes with respect to usefulness. We are in the process of finalising a distribution of the WRS and we hope that there will be enough interest in our prototype to provide us with empirical evidence to answer both of these questions.

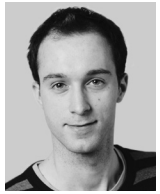
References

1) Wikipedia: size of wikipedia, http://en.wikipedia.org/wiki/Size_of_wikipedia (Nov. 2010).
2) Adler, B.T., Chatterjee, K., de Alfaro, L., Faella, M., Pye, I. and Raman, V.: Assigning trust to wikipedia content, *Proc. 4th International Symposium on Wikis* (*WikiSym'08*), Porto, Portugal (Sep. 2008).
3) Adler, B.T. and de Alfaro, L.: A content-driven reputation system for the wikipedia, *Proc. 16th International World Wide Web Conference*, Banff, Alberta, Canada, pp.261–270 (May 2007).
4) Citizendium: The Citizen's Compendium: Welcome to citizendium, http://en.citizendium.org/wiki/Welcome_to_Citizendium (Jan. 2010).
5) Library of Congress: Library of congress classification outline, http://www.loc.gov/catdir/cpso/lcco/ (visited 4 Jan. 2010).
6) Dondio, P., Barrett, S., Weber, S. and Seigneur, J.M.: Extracting trust from domain analysis: A case study on the wikipedia project, *ATC*, pp.362–373 (2006).
7) Goldberg, D., Nichols, D., Oki, B.M. and Terry, D.: Using collaborative filtering to weave an information tapestry, *Comm. ACM*, Vol.35, No.12, pp.61–70 (1992).
8) Jensen, C.D.: *Building a reputation system for the Wikipedia*, chap. Open Innovation, Göttingen University Press (2011), to appear.
9) Jensen, C.D. and Korsgaard, T.R.: Dynamics of trust evolution: Autoconfiguration of dispositional trust dynamics, *Proc. International Conference on Security and Cryptography* (*SECRYPT 2008*), Porto, Portugal, pp.509–517 (July 2008).
10) Jensen, C.D. and Lefévre, T.: Evaluating the expertise of recommenders in the wikipedia recommender system, *Short Paper Proc. IFIPTM 2010*, Morioka, Iwate, Japan (June 2010).
11) Korsgaard, T.R.: Improving trust in the Wikipedia, Master's thesis, Technical University of Denmark, Department of Informatics & Mathematical Modelling (2007).
12) Korsgaard, T.R. and Jensen, C.D.: Reengineering the Wikipedia for reputation, *Electronic Notes in Theoretical Computer Science*, Vol.244, pp.81–94 (Aug. 2009).
13) Lefévre, T.: Extending the Wikipedia Recommender System: Assessing Expertise of Recommenders, Master's thesis, Technical University of Denmark, Department of Informatics & Mathematical Modelling (2009).
14) Lefévre, T., Jensen, C.D. and Korsgaard, T.R.: WRS: The Wikipedia Recommender System, *Proc. 3rd IFIP WG 11.11 International Conference* (*IFIPTM 2009*), Purdue University, West Lafayette, Indiana, U.S.A., pp.298–302 (June 2009).
15) Online Computer Library Center (OCLC): Introduction to the Dewey Decimal Classification, http://www.oclc.org/dewey/versions/ddc22print/intro.pdf (downloaded 6 Jan. 2010).
16) Open Directory Project: About the Open Directory Project, http://www.dmoz.org/docs/en/about.html (visited 3 Nov. 2010).
17) Orlowski, A.: Avoid wikipedia, warns wikipedia chief, it can seriously damage your grades, The Register (June 2006).
18) Pilkauskas, P.: Expertise classification of recommenders in the Wikipedia Recommender System, Master's thesis, Technical University of Denmark, Department of Informatics & Mathematical Modelling (2010).
19) Sanger, L.: Why Wikipedia must jettison its anti-elitism, Kuro5hin.org: Op-Ed (Dec. 2004), http://www.kuro5hin.org/story/2004/12/30/142458/25
20) Schein, A.I., Popescul, A., Ungar, L.H. and Pennock, D.M.: Methods and metrics for cold-start recommendations, *Proc. 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR '02*), Tampere, Finland, pp.253–260 (2002).
21) Seigenthaler, J.: A false Wikipedia 'biography', *USA TODAY* (Nov. 2005).
22) Suh, B., Chi, E.H., Kittur, A. and Pendleton, B.A.: Lifting the veil: Improving accountability and social transparency in wikipedia with wikidashboard, *Proc. CHI 2008*, Florence, Italy, pp.1037–1040 (Apr. 2008).
23) UDC Consortium: UDC Summary, http://www.udcc.org/udcsummary/php/index.php (visited 4 Jan. 2010).
24) Viégas, F.B., Wattenberg, M. and Dave, K.: Studying cooperation and conflict between authors with *history flow* visualizations, *Proc. CHI 2004*, Vienna, Austria, pp.575–582 (Apr. 2004).

(Received November 19, 2010)
(Accepted January 14, 2011)

(Original version of this article can be found in the Journal of Information Processing Vol.19, pp.345–363.)

Christian Damsgaard Jensen holds an M.Sc. in Computer Science from the University of Copenhagen (Denmark), a Ph.D. in computer science from Université Joseph Fourier (Grenoble, France) and an M.A. (jure officii) from Trinity College Dublin (Ireland). He is an associate professor at the Department of Informatics and Mathematical Modelling at the Technical University of Denmark, where he teaches and conducts research in the area of security in open distributed systems. For the past 10 years, he has focused on trust-based methods and technologies to secure collaboration among entities in open distributed system. This work addresses all 3 As in AAA: Authentication technologies and entity recognition; Access control policies and mechanisms; and Accountability through reputation and recommendation systems.



Povilas Pilkauskas graduated from the Technical University of Denmark in October 2010 with a M.Sc. in Engineering. His thesis entitled Expertise classification of recommenders in the Wikipedia Recommender System, which investigated different information classification schemes for classifying articles in the Wikipedia. He is focused on new Web technologies impacts for current Web society including end-users and developers. Povilas Pilkauskas is still contributing to the Wikipedia Recommender System Project.

Thomas Lefévre graduated from the Technical University of Denmark in 2009 with an M.Sc. in Engineering. Since his graduation, he has worked for the company Traen A/S as a consultantn, developing custom web applications for clients, such as the Danish Ministry of Education & The Danish Medicines Agency. In his spare time, he still contributes to the Wikipedia Recommender System Project.