Automatic Knowledge Retrieval from the Web

Marcin Skowron and Kenji Araki

Graduate School of Information Science and Technology, Hokkaido University,
Kita-ku Kita 14-jo Nishi 8-chome, 060–0814 Sapporo, Japan

Abstract. This paper presents the method of automatic knowledge retrieval from the web. The aim of the system that implements it, is to automatically create entries to a knowledge database, similar to the ones that are being provided by the volunteer contributors. As only a small fraction of the statements accessible on the web can be treated as valid knowledge concepts we considered the method for their filtering and verification, based on the similarity measurements with the concepts found in the manually created knowledge database. The results demonstrate that the system can retrieve valid knowledge concepts both for topics that are described in the manually created database, as well as the ones that are not covered there.

1 Introduction

Despite the years of research in the field of Artificial Intelligence, the creation of a machine with the ability to think is still far from realization. Although computer systems are capable of performing several complicated tasks that require human beings to extensively use their thinking capabilities, machines still cannot engage into really meaningful conversation or understand what people talk about. One of the main unresolved problems is the lack of machine usable knowledge. Without it, machines cannot reason about the everyday world in a similar way to human beings. In the last decade we have witnessed a few attempts to create knowledge databases using various approaches: manual, machine learning and mass collaboration of volunteer contributors. So far, none of these approaches can be considered as fully reliable and/or efficient. To a various extent, each of the presented methods has some problem areas and limitations that are inherent to it. We discuss this issue in more detail in the next section.

Compared to the efforts taken to create knowledge databases manually, either as an effort of one entity or distributed among several parties, surprisingly little has been done to develop the methods for the automatic retrieval of knowledge concepts for such databases. This paper presents a language independent method, capable to retrieve general and commonsensical knowledge for the open-domain applications from the web and its implementation to the working system. The aim of the system that implements it, is to automatically create entries to a knowledge database, similar to the ones that are being provided by the volunteer contributors. The system uses Internet as a source of knowledge concepts candidates. Obviously, only a small fraction of the statements that are accessible on the web can be considered as

a valid entries to a knowledge database. Below, we describe the methods for "web knowledge concepts" filtering and verification, which are based on the similarity comparison with concepts found in a knowledge database created with the mass collaboration of Internet users. The paper also presents the implementation of these methods to the developed system and preliminary results obtained from it.

2 Approaches to Knowledge Database Construction

The most well known example of the manual approach to knowledge base construction is the CYC project, which contains 1.5 million assertions build over 15 years[4]. The aim of the project was to create a database that could provide knowledge from a large scope of domains, along with the means to effectively use this knowledge for systems that could engage in the reasoning of human affairs. We learn from the experiences of this and similar projects that building a database in this way was laborious, time-consuming and costly. We argue that in an open domain where new information appears and becomes obsolete on a daily basis, a complete knowledge-base build using this approach is out of reach.

The machine learning approach demonstrated that it was feasible to automatically induce rules from data and to overcome some problems characteristic for the manual approach presented above. However, as of yet the machine learning approach has not resulted in creation of a large, open-domain knowledge base. A typical learning program has only weak assumptions about the world; consequently, the learned rules are relatively shallow as they refer only to correlations between observable variables[7]. To address this problem the researchers attempted to incorporate pre-existing knowledge, combining the automatic and manual approaches. However, they soon faced the bottleneck related to collecting knowledge known from the manual approach.

Knowledge acquisition from Internet users is the approach used in the construction of the Open Mind Common Sense database (OMCS)[10], allowing mass collaborations of thousands of volunteer contributors. The ability to input knowledge concepts directly in the natural-language form simplified the process of database building compared to the attempts that used semi-programming languages. Thanks to this, theoretically every (English speaking) Internet user, without any special training can input knowledge concepts using plain sentences. As demonstrated by Borchardt[2] and Singh[10] such concepts can be used for reasoning. While significantly decreasing the amount of money spent, this approach still requires large investment of human labor distributed among thousands of Internet users. Compared to the manual approach, the time requirement for creation of the knowledge database is reduced; but, this factor cannot be overlooked. The challenges frequently faced in the projects that use the mass collaboration approach include the need to ensure[7]:

- High quality of contributed knowledge,
- Consistency between knowledge entered by different contributors and at different times,
- Relevance of inputted knowledge to a given task,
- Scalability of the project,
- Motivation of contributors to start and consistently work on the project.

So far the OMCS project has gathered more than 700,000 items from more than 15,000 users. The database was evaluated[10] by human judges using a sample of the knowledge concepts, and the following conclusions were presented: 75% of the items are largely true, 82% are largely objective, 85% were judged as largely making sense. Tables 1 and 2 show the examples of knowledge concepts related to the nouns "water" and "apple" from the OMCS database.

Table 1. Examples of knowledge concepts related to "water" from the OMCS.

No.	**Knowledge concept related to "water"**
1	The last thing you do when you take a shower is turn off the water
2	Human beings need water to survive
3	When animals need water, they feel thirsty
4	Clouds are made up of water vapor
5	People need to drink water verey day

Table 2. Examples of knowledge concepts related to "apple" from the OMCS.

No.	**Knowledge concept related to "apple"**
1	Yellow apples are soft and sweet
2	The first thing you do when you eat an apple is rub it
3	An apple contains seeds
4	The Michigan state flower is the apple blossom
5	When you drop an apple, it gets bruised

3 Automatic Knowledge Retrieval from the Web

Analyzing the content of knowledge database created by volunteer contributors, one can discover that several of the statements found there exist also on freely available web pages. Additionally, the number of "web knowledge concepts", using slightly different words and/or syntax, semantically provides equivalents for a large part of the entries from the manually constructed knowledge-bases. Other knowledge concepts accessible on the web describe

topics that are not covered yet in manually created databases or provide wider coverage and additional details for several of the concepts defined in the manually created database.

We argue that the web is a rich resource of commonsensical and general knowledge and that this resource is usable in the process of automatic creation of knowledge databases. For automatic knowledge retrieval the web has important advantages, including real-time updates of content, wide coverage of various domains, and diversity of presented opinions. At present, a popular search engine indexes more than $8*10^9$ web pages. Assuming only a small portion of them include statements that can be treated as valid entries to a knowledge database, the web still hosts an immense number of knowledge concepts that can be automatically retrieved. We think that in every moment, in various part of the world Internet users/WWW creators contribute the knowledge that can and ought to be used for the building of the knowledge databases and supporting several AI applications. Obviously, there is a need to filter out statements that cannot be considered as reliable and valid entries to a knowledge database. The main challenge in this approach is to ensure a high recall rate of knowledge concepts from various domains and precision of concepts filtering and selection.

4 Relevant Research

Some relevant research adapting different approaches, and to some extent having different aims, is included in the following works. Studies on knowledge representation and classification were done by Woods, followed by work on automatic construction of taxonomies by extracting concepts from texts[11]. Satoh [8] use connective markers to acquire casual knowledge similarly to the later work of Inui [3], where the method for classification of casual relations was presented. The research of Rzepka[9] described the methods for automatic word categorization and script generation using the statistic of words obtained from a Web-sized corpus and by exploring Japanese grammar features. The work of Alani[1] presented a method of automatic knowledge extraction from a selected domain (knowledge about artists) from web documents based on a predefined ontology. In the most closely related work[6], Moldovan described a system that gathers knowledge from a financial domain. However, this system was tailored to a specific domain and its processing could not be done automatically, as the system required an interaction with a user to verify the acquired concepts, relationships and extraction patterns.

5 System Processing

Below we present the "KnowY" system that implements the idea of automatic retrieval of open-domain knowledge concepts from the web. There are two aims of this system: to automatically create a knowledge database similar to

ones that are being built manually through the mass collaboration of Internet users, and to support various AI applications with the knowledge required in their processing. At present, the system utilizes the OMCS database to learn what constitutes a valid entry to a knowledge database. This information is necessary for filtering out statements that are unlikely to form such an entry and to rank discovered concepts depending on their similarity to ones found in the OMCS database. The adaptation to any other knowledge database written in natural language form is also feasible. The system is developed using Perl and C, and implemented on the single Pentium III-class Linux 2.4 workstation. For locating and accessing knowledge concepts on the web "KnowY" uses Yahoo search engine[1]. Figure 1 presents the system flowchart.

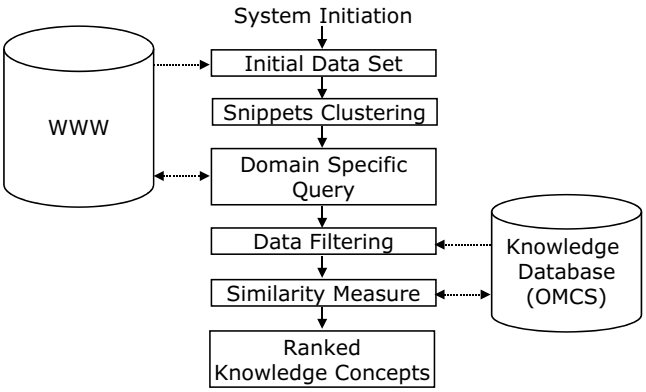

Fig. 1. System flowchart

"KnowY" is initiated by receiving a request for a noun or list of nouns for which knowledge concepts are to be discovered. To provide a wide representation of various topics, the system retrieves the 300 highest scored snippets for each of the 5 different domains (.gov, .edu, .net, .org, and .com) using a submitted noun as a query. A set of the unique snippets is clustered[5], and for each cluster the 3 most descriptive words are discovered. Initially the number of clusters is set to 4. For example for the noun "apple" the following most descriptive words characterizing the content of the clusters have been found (tree, fruit, apples), (juice, production, cider), (computer, mac, macintosh), (iTunes, music, server).

In the next step, "KnowY" executes the new search using a query consisting of the requested noun along with the 3 most descriptive words discovered for each of the discovered clusters. The system retrieves a set of 300 web pages, which provides a wide selection of information concerning the noun. Extension of a used query with the 3 most descriptive words for a given clus-

[1] http://www.yahoo.com

ter ensures that domain limitations are respected. After dividing the text into sentences, the ones that do not include the requested noun (either in singular or plural form) are excluded.

In the data filtering and similarity measure stages, the OMCS database was used as training data that revealed the most frequent ways of describing knowledge concepts, as well as commonly used words and grammatical constructions[2]. The comparison set is composed of the concepts that include the requested noun (either in singular or plural form). If the number of such concepts is less than 100, the system randomly adds additional concepts to obtain such a set. As our experiments show, for a well defined concept (high quality and wide coverage of a given noun in the OMCS in the database), the former method is likely to find a slightly better set of automatically retrieved "web knowledge concepts". On the other hand, the later approach provides the means to discover new concepts, and allows "KnowY" to generate a database for terms that are not covered by the OMCS at all. We think that this feature is of prime importance for many applications in the open domain. The average sentence length found in the OMCS database was 9.32 words long. For the similarity score ranking we decided to include only "web sentences" and OMCS concepts with the number of words between 3 and 20. In the filtering process "KnowY" also uses information on the proportion of alphanumeric and special characters. To exclude the sentences that are unlikely to be valid entries to the knowledge database, "KnowY" compares them to the OMCS knowledge database concepts and ranks using the highest similarity score obtained, calculated with the following formula:

$$Similarity = \frac{\sum_{n=1}^{N} Wn * Wn_{ITF}}{L1 + L2} + \log Np * \alpha \tag{1}$$

,where: Wn - matching word found both in a OMCS concept and a "web sentence", Wn_{ITF} - inverted term frequency for a matching word (OMCS), L1 - number of words in a "web sentence", L2 - number of words in a OMCS concept, Np - number of concepts from OMCS where a noun was found in a position from a "web sentence", α - parameter (the value of α was set to 0.2). The formula takes account of the number of matching words between a "web sentence" and an OMCS concept as well as the importance of the given word in a used set of OMCS terms by the means of the ITF. The Np value promotes "web sentences", where a noun appears on a position, which is frequent for many OMCS concepts.

[2] In our experiments we used the snapshot of the OMCS database consisting of 700,000 sentences, available at http://commonsense.media.mit.edu/cgi-bin/download.cgi. From this set the sentences describing pictures and stories were removed.

6 Experiment Results

The preliminary experiments with the system were performed using a set of nouns, including ones that are frequently described in OMCS, as well as ones that are not covered at all in this database. Tables 3 and 4 show the examples of the knowledge concepts automatically discovered from the web; only the first five highest ranked statements are presented.

Table 3. Highest ranked knowledge concepts discovered by "KnowY" related to "water", domain (drinking, epa, treatment).

No.	**Discovered Knowledge Concept**	**Sim. score**
1	Sanitation means not only clean water but also clean air and clean soil.	1.33
2	Clouds are made when water vapor condenses into tiny droplets.	1.29
3	Most lakes are filled with fresh water, but there are a few lakes that are filled with salt.	1.25
4	When animals need water, they should not have to stand and wait.	1.17
5	Overwatering your yard can also cause water to run into the streets and into storm drains.	1.12

Table 4. Highest ranked knowledge concepts discovered by "KnowY" related to "apple", domain (tree, fruit, apples).

No.	**Discovered Knowledge Concept**	**Sim. score**
1	Apples bruise easily and must be hand picked.	1.61
2	Apple blossom is the state flower of Michigan.	1.58
3	Provide small samples different types of apples.	1.28
4	Apples have 5 seeds.	1.22
5	Apples are a member of the rose family.	1.21

As the results demonstrate, "KnowY" is able to automatically retrieve the knowledge concepts respecting the domain limitations with relatively high accuracy. Furthermore, the experiments showed that the system is capable of finding and automatically retrieving from the web the semantic equivalents of several of the entries that were inputted manually to the OMCS database. An example of such a statement discovered for the noun "apple", which is present among first five concepts with the highest similarity score is "Apple blossom is the state flower of Michigan", and its counterpart from the OMCS, "The Michigan state flower is the apple blossom". Some of the knowledge concepts

obtained from the web provide more detail compared to the OMCS entries. The instances of such "web knowledge concepts" include "Clouds are made when water vapor condenses into tiny droplets", " Apples bruise easily and must be hand picked", " Apples have 5 seeds" and the corresponding OMCS database entries, " Clouds are made up of water vapor"," When you drop an apple, it gets bruised", and " An apple contains seeds". With the exception of the statement " Provide small samples different types of apples"(Table 4, pos 3.) all remaining, automatically retrieved statements can be treated as valid entries to a knowledge database.

Less reliable set of results was obtained for the noun "golf" in the domain "clubs, instruction, equipment". As shown in Table 5, the majority of the discovered sentences can not be considered as valid knowledge concept since they convey personal experience/opinion (pos. 2), or strictly commercial information (pos. 4 and 5).

Table 5. Highest ranked knowledge concepts discovered by "KnowY" related to "golf", domain (clubs, instruction, equipment).

No.	**Discovered Knowledge Concept**	**Sim. score**
1	Golf is a great game that you can play for a lifetime...enjoy it!	1.02
2	The men I usually play golf with refused to play golf with me until I stepped back.	0.99
3	Good golf techniques are simple to learn but must be reinforced to be effective.	0.94
4	Over 600 courses, 200 hotels, and 50 plus golf resorts to choose from!	0.93
5	Golf for Beginners will provide you with the necessary help you need to get started correctly.	0.93

Table 6 and 7 present the results obtained for the noun "wasabi" and "Kendo". These nouns are covered only to a very limited extend in the OMCS ("wasabi" - 7 entries) or do not appear in this database ("Kendo"). For the similarity score calculation the comparison set included 93 and 100 randomly selected statements from the OMCS database, respectively for the nouns "wasabi" and "Kendo". The average similarity score is considerably lower, compared with the score obtained for the automatically discovered knowledge concept related to "water" or "apple". However, although the comparison set was composed of mostly randomly selected concepts from OMCS, in our opinion it did not significantly compromise the quality of the discovered concepts. The majority of retrieved statements related to "wasabi" and "Kendo" could be included in a knowledge database. The exception is sentence " This being the modern day Kendo's source of philosophy"(Table 7, pos. 3), which according to the standards used in the evaluation of the OMCS database would not be classified as a complete and valid knowledge concept.

Table 6. Highest ranked knowledge concepts discovered by "KnowY" related to "wasabi", domain (sushi, japanese, horseradish).

No.	Discovered Knowledge Concept	Sim. score
1	That is one reason that wasabi is served with sushi and raw fish slices.	0.91
2	Wasabi is very very dangerous.	0.86
3	Wasabi A green pungent horseradish paste served with sushi and sashimi.	0.82
4	What sushi lovers are unaware of however, is that Wasabi served in America is seldom real.	0.72
5	In Japan, sushi and sashimi are served with a condiment of wasabi mixed with soy sauce.	0.70

Table 7. Highest ranked knowledge concepts discovered by "KnowY" related to "Kendo", domain (sword, arm, Japanese).

No.	Discovered Knowledge Concept	Sim. score
1	Anger and true aggression has nothing to do with Kendo.	0.73
2	The bokken is used in modern kendo for kata practice.	0.67
3	This being the modern day Kendo's source of philosophy.	0.65
4	There are eight striking points in Kendo used for scoring.	0.62
5	Shiai geiko is the most competitive part of Kendo.	0.61

7 Conclusions and Future Work

This paper presented the method of automatic knowledge retrieval from the web. We think the idea described here and implemented in "KnowY" contributes to address one of the most important challenges in AI that exists today. The experiments performed with the system demonstrated that it is capable of automatically discovering several knowledge concepts in a user-selected domain with relatively high accuracy. Some of the automatically retrieved knowledge concepts provided semantic equivalents of the statements that were manually inputted to the OMCS. Others, while including more details compared to the OMCS entries, could also become a part of a knowledge database. The results confirmed also that the system was able to retrieve high quality knowledge concepts, even for the terms that were not described in the knowledge database built by mass collaboration of Internet users.

In our future work we are focusing on evaluating other techniques proposed for the document similarity in the information retrieval and computational linguistics literature; especially those that could be effectively applied for the comparison of short, sentence long documents. Some shortcomings of the described method are related to the nature of the resource that it is based on; while the web provides wide coverage of various domains, very

frequently the amount of commercial information dominates over these that could be considered as valid knowledge concepts. To address this problem, we intend to extend queries used to access "web knowledge concepts" with a short list of negative words (frequently used on the pages that contain advertisements), which should not appear on the pages "KnowY" accesses, as well as to extend the similarity measurement formula with the means to penalize the statements that include strictly personal opinions or information of only a commercial nature. Given that the described method is language independent, we intend also to evaluate it using languages other than English.

References

1. Alani H., Kim S., Millard D., Weal M., Lewis P., Hall W. and Shadbolt N. Automatic Ontology-based Knowledge Extraction and Tailored Biography Generation from the Web. *IEEE Intelligent Systems, 18(1)*, pages 14-21, 2003.
2. Borchardt G. Understanding Casual Descriptions of Physical Systems. In Proceedings of the Tenth National Conference on Artificial Intelligence, pages 2-8, 1992.
3. Inui T., Inui K. and Matsumoto Y. What Kind and Amount of Casual Knowledge Can Be Acquired from Text by Using Connective Markers as Clues? In the 6th International Conference on Discovery Science, pages 179-192, 2003.
4. Lenat D. CYC: A large-scale investment in knowledge infrastructure. Communications of the ACM, 38(11), pages 33-38, 1995.
5. Karypis G. A Clustering Toolkit. http://www.cs.umn.edu/~karypis/cluto. 2003.
6. Moldovan D., Girju R. and Rus V. Domain-Specific Knowledge Acquisition from Text. Proceedings of the Applied Natural Language Processing Conference, pages 268-275, 2000.
7. Richardson M. Domingos P. Building Large Knowledge Bases by Mass Collaboration. Proceedings of the Second International Conference on Knowledge Capture, pages 129-137, 2003.
8. Satoh H. Retrieval of simplified casual knowledge in text and its applications.
9. Rzepka R., Itoh T. and Araki K. Rethinking Plans and Scripts Realization in the Age of Web-mining. IPSJ SIG Technical Report 2004-NL-162, pages 11-18, 2004.
10. Singh P. The public acquisition of commonsense knowledge. In Proceedings of AAAI Spring Symposium: Acquiring (and Using) Linguistic (and World) Knowledge for Information Access, 2002.
11. Woods W. A Better way to Organize Knowledge. Technical Report of Sun Microsystems Inc., 1997.