A Survey of Crowdsourced Knowledge Acquisition

Rui Meng

Supervisor: Lei Chen

Department of Computer Science and Engineering

Hong Kong University of Science and Technology

29th Nov., 2016

Abstract

Knowledge acquisition, which refers to the process of extracting, structuring and organizing knowledge from some sources, empowers the computer programs to reason and solve complex problems. Recent years have seen significant advances in the field of knowledge acquisition. While notable efforts have been devoted, the state-of-the-art automatic techniques for knowledge acquisition still have its limitations and can yield noisy or semantically meaningless knowledge. Crowdsourcing, which has gained its popularity in various domains to handle hard tasks with human involvement, offers an alternate approach for knowledge acquisition. Crowdsourcing and human computing, on the one hand, can improve the information extraction technique from text inputs to acquire accurate and clean knowledge; on the other hand, as a natural source of knowledge, the crowd can be mined to obtain knowledge that resides in the human mind, such as commonsense knowledge, individual knowledge and subjective knowledge. With the large scale of knowledge and limited crowdsourcing resource, hybrid systems have been proposed to combine the power of machine-based automatic technique and crowdsourcing approach for knowledge acquisition.

Throughout this survey, we give a comprehensive overview of crowdsourced knowledge acquisition techniques. We first clarify the key concepts and relevant technical backgrounds of automatic knowledge acquisition. Then, we present the crowdsourced knowledge acquisition approaches and the hybrid frameworks that combine the machine-based automatic approach and crowdsourcing technique. Finally, we identify some future research issues in the crowdsourced knowledge acquisition area.

Contents

1 Introduction

Knowledge Acquisition (KA) is the process of acquiring (either directly from human experts, books , documents, sensors, or computer files) information and its formalized structure for a knowledge-based system. The derived structured information will allow some particular task to be performed by a computer system. Throughout this survey, structured information of this sort is commonly termed "knowledge". Knowledge acquisition typically begins with the process of receiving or acquiring new information such as information extraction technique; once information is received, knowledge acquisition continues with information encoding and understanding to build schema, i.e. ontology, knowledge base, taxonomy, etc. Then, knowledge acquisition also include the process of recall and alter the stored information. Recently, the knowledge acquisition has gained its popularity with the creation of large-scale knowledge bases, such as YAGO [38], DBpedia [6], Freebase [7], DeepDive [31], NELL [11], KnowItAll [15]. These large-scale knowledge bases organize the structured information as knowledge fact, typically as the subject-predicate-object triples consisting of entities and binary relationship among entities. The KBs empowers various applications such as short-term understanding, query refinement, machine-translation, question answering and so forth, therefore, advanced techniques have been developed for knowledge acquisition and knowledge base construction.

Considering the information source, there are mainly two types: text information and human knowledge. Text information covers web pages, web tables, books documents and so forth, these sources contains large-scale information and potential valuable knowledge. Therefore, automatic knowledge acquisition techniques have been proposed to extract information and knowledge [41, 16, 43, 31]. The two major tasks are information extraction, how to extract entities and relations from the vast amount of raw text, and knowledge base construction, how to encode the extracted entities and relations into machine-interpretable knowledge facts. While advanced techniques have been developed, the automatic approach rely on machine intelligence suffer from the low accuracy results. On the one hand, the information source are not clean enough which may contain out-of-date data, in-consistency data

or wrong data; on the other hand, the semantic ambiguity resides in the information sources as there are various expressions for a single fact. Moreover, the knowledge that expressed in the text information sources is called explicit knowledge which is objective. Another kind of knowledge is called tacit knowledge, which is subjective and experiential, this kind of knowledge resides in human mind and can not be acquired by machine-based techniques and text information inputs.

Crowdsourcing and human computation are emerging fields that sit squarely at the intersection of economics and computer science. They examine how people can be used to solve complex tasks that are currently beyond the capabilities of artificial intelligence algorithms. Online marketplaces like Mechanical Turk [1] and CrowdFlower [2] provide an infrastructure that allows micro-payments to be given to people in return for completing human intelligence tasks (HITs). Recently, crowdsourcing and human computation have gained its popularity and been successfully employed in various domains to handle human intrinsic tasks and hard tasks where fully automatic solutions are deemed inadequate, such as data cleaning [14], table matching [17]. As human is a natural source of tacit knowledge and its capability of knowledge understanding, crowdsourcing and human computation is an alternative to overcome the limitations of machine-based automatic knowledge acquisition. Some works have been tried to leverage the power of crowdsourcing for knowledge acquisition and knowledge base construction [37, 33, 31, 18, 38]. However, as the micro-payments have to be given for HITs, the crowdsourcing technique is usually limited by the monetary budget. Therefore, although have a boarder knowledge coverage and better accuracy in knowledge extraction, the crowdsourcing itself can not handle large scale knowledge acquisition tasks.

Motivated by the aforementioned shortcoming of automatic machine-based techniques and crowdsourcing techniques for knowledge acquisition, some hybrid frameworks [21, 29] have been proposed to combine the power of machine-based information extraction and KB construction and crowdsourcing to overcome the limitations, which can improve the knowledge acquisition quality and lowers the crowdsourcing task. The hybrid systems concerns two major sub-tasks: how to effectively integrate the machine-based approach with the crowd-

sourcing and how to design HITs in the crowdsourcing stage to lower the cost and insure the quality.

To sum up, in this article, we conduct a comprehensive survey of crowdsourced knowledge acquisition techniques. We first review the state-of-the-art automatic knowledge acquisition technique, knowledge acquisition approach relying on crowdsourcing and the hybrid approach that combine the power of machine and crowdsourcing. Then, we propose our view of crowdsourced knowledge acquisition and identify some future research issues in this area.

2 Background

2.1 Knowledge Representation

Knowledge representation targets at organizing the acquired knowledge so that it can be ready for use, i.e. computer programs to perform searching and reasoning. A knowledge base (KB) is a technology used to encode the structured and semi-structured information, knowns as knowledge facts, used by a computer system. In KBs, the entities and the binary relationships among entities are stored as logical assertions in the form of subject-predicate-object triples.

A ***knowledge base (KB)*** is a tuple denoted by (E, L, R, P), consisting of a collection of **entities** E, **literals** L, **relations** R holding between entities and **properties** P holding between entities and literals. Each **entity** $e \in E$ is unique in each KB. An entity can be a class/concepts or an instance, i.e. $E = \{C \cup I\}$, where C and I represent a class set and an instance set. An instance corresponds to a real world object, and a class is a set of instances that share common properties. Note that in a KB, the mapping from an instance to a class can be found through "type_of" relationship. A **knowledge fact** is a subject-predicate-object triple $< s, p, o >$, where s is an entity in a KB, p is a relationship or property in a KB and o is either an entity in a KB or literals. Therefore, a KB consists of a collection of *entities, literals, relations and properties* and the knowledge is encoded in the form of *knowledge facts*. Figure 1 shows a toy example of KB, there are four entities - two classes, "politician" and "president" and two instances, "Obama" and "Michelle"; the date "1961-8-4" and string "Barack Obama" are literals; three relations "subclass", "type_of" and "married_to" and two properties "born" and "full_name". There are five knowledge facts in the toy example, they are: <Obama,type_of,president>, <president,subclass,politician>,<Obama,married_to,Michelle>,<Obama,born,1961-8-4> and <Obama,full_name,Barack Obama>.

Among the models that encode knowledge bases, the Resource Description Framework (RDF) model is the most popular one. RDFS (RDF Schema) extends RDF to incorporate type information. Web Ontology Language (OWL) extends RDFS by providing terminology.

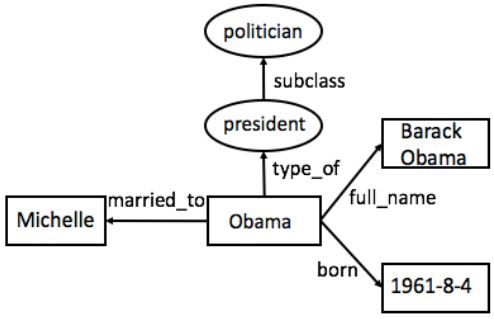


Figure 1: An example of knowledge base

Knowledge base systems can extend the RDFS/OWL model based on its requirements. SPARQL is adopted to perform queries and reasoning in KB system.

2.2 Crowdsourcing and Human Computing

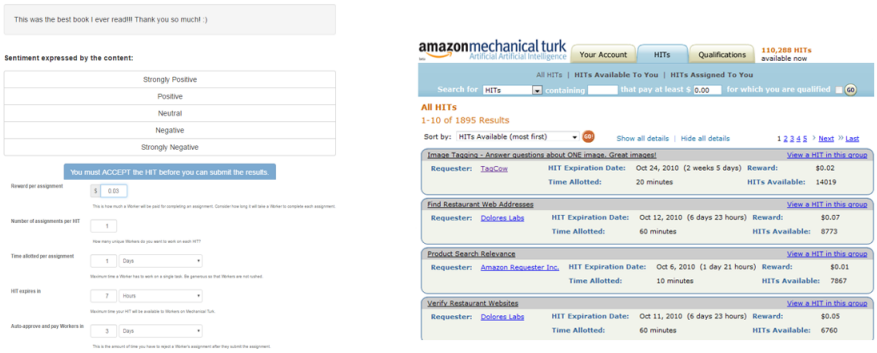
Human computation has been used for centuries, which describes computation performed by a human and human computation systems organize human efforts to carry out computation [26]. large crowd of users to solving computationally hard problems. The idea is to leverage the innate cognitive and intellectual capabilities of humans for certain tasks, especially those human-intrinsic tasks that the machine-based solutions has poor performance, such as entity resolution [40, 17, 39], table matching [17], schema matching [44], information acquisition [22], taxonomy construction [12, 8], ontology integration [34], data cleaning [14] and so forth.. With the popularity of commercial crowdsourcing platforms such as Amazon Mechanical Turk (AMT) [1] and CrowdFlower [2], the source of human is broadened to the crowd, namely Crowdsourcing. Recently, the increasing popularity of crowdsourcing brings new trends to adopt online crowd as a Human Processing Unit (HPU) and leverage the power of the crowd to tackle human intrinsic tasks. Many tasks that can not be addressed by machines perfectly have been solved or improved by crowdsourcing techniques.

The three key crowdsourcing genres [32]:

- Mechanised labor (MLab) is a type of paid-for crowdsourcing, where contributors choose to carry out micro-tasks and are paid a small amount of money in return (often referred to as micro-payments). The most popular platform for mechanised labor is Amazon Mechanical Turk (MTurk) which allows requesters to post their micro-tasks in the form of Human Intelligence Tasks (or HITs) to a large population of micro-workers (often referred to as turkers). There are several key concepts of MTurk: **Requester**. A Requester is a company, organization, or person that creates and submits tasks to crowdsourcing platforms for Workers to perform. As a Requester, you can use a software application to interact with the platform to submit tasks, retrieve results, and perform other tasks. **Worker**. A Worker is a person who performs the tasks specified by a Requester in a task. Workers can find and accept assignments, enter values into the question form, and submit the results. The Requester specifies how many Workers can work on a task. The crowdsourcing platform guarantees that a Worker can work on each task only one time. **Human Intelligence Task (HIT)**. A Human Intelligence Task (HIT) is a task that a Requester submits to the crowdsourcing platform for Workers to perform. A HIT also has an assignment duration, which is the amount of time a Worker has to complete a HIT after accepting it. **Assignment**. An assignment specifies how many people can submit completed work for your HIT. When a Worker accepts a HIT, the platform creates an assignment to track the work to completion. **Reward**. A reward is the money the requester pay Workers for satisfactory work they do on HITs. Figure 2 illustrated the aforementioned concepts using the Amazon Mechanical Turk. The HIT creation interface and HIT answer interface are shown.

- Games with a purpose (GWAPs) enable human contributors to carry out computation tasks as a side effect of playing online games. The challenges in using GWAPs in scientific context are in designing appealing games and attracting a critical mass of players.

- Altruistic crowdsourcing refers to cases where a task is carried out by a large number of volunteer contributors. To reduce the incentive to cheat (e.g., for money or glory),



HIT Creation by Requester HIT Answer by Worker

Figure 2: HITs on Amazon Mechanical Turk

altruistic crowdsourcing approaches leverage the intrinsic motivation of a community interested in a domain.

As each HIT is associated with a monetary reward in the MLab genre, the crowdsourcing technique usually suffer from the trade-off among answer quality, answer latency and monetary cost. Moreover, human are error-prone, as studied, the quality of normal workers is around 75% even with the simple tasks. Therefore, in order to lower the crowdsourcing task, reduce the latency and improve the answer quality, much work have been conducted on the crowdsourcing mechanisms including task assignment, task pricing and quality control.

3 Knowledge Acquisition by Information Extraction

In this section, we review some state-of-the-art information extraction methods for knowledge acquisition. Conventional information extraction methods try to extract knowledge from texts with natural language inputs. We introduce information extraction technique for relation and fact extraction and the approaches for knowledge base construction.

3.1 Information Extraction Technique

Information extraction (IE) refers to the process of identifying the instances of facts (names, entities, relations and events) from semi-structured or unstructured text; and convert them into structured representations. Their underlying methods are based on rules and patterns, linguistic analysis, statistical learning and often combinations of all these elements. Traditional methods focused on harvesting tuples that satisfy prespecified relations from a domain, while the more recent OpenIE systems are data-driven, aiming to capture all relational tuples from heterogeneous sources. We briefly discuss some of these methods below.

Pattern based IE. Pattern-based approaches adopt syntactic patterns to extract and discover relationships. Pattern-based approaches have high accuracy if the patterns are carefully chosen. Authors [18] adopt high quality syntactic patterns, (*"Hearst Patterns"*, shown in Figure 3) to extract the "isA" relationships. However, these approaches suffer from the sparse coverage problem since high quality patterns are rare [41]. A. Ritter [33] and R. Snow [37] try to improve the recall and coverage by exploring coordinate relations to learn more potential "isA" patterns. DIPRE [10] and Snowball [3] try to learn patterns implicitly in an iterative process: starting with a bootstrap tuple-set of the target relation, occurrences of all tuples in the corpus are extracted along with their context. In the next step, patterns are generated based on the tuple occurrences and their surrounding contexts. Finally, the sample is expanded with the newly found patterns and searched again.

Linguistic analysis based IE. Some work use the linguistic features for relation extraction tasks. There are two kinds of linguistic analysis: deep linguistic analysis which provides

ID	Pattern
1	*NP* such as {*NP*,}* {(or | and)} *NP*
2	such *NP* as {*NP*,}* {(or | and)} *NP*
3	*NP*{,} including {*NP*,}* {(or | and)} *NP*
4	*NP*{,*NP*}*{,} and other *NP*
5	*NP*{,*NP*}*{,} or other *NP*
6	*NP*{,} especially {*NP*,}* {(or | and)} *NP*

Figure 3: Hearst Pattern

a syntactic parsing of each sentences and shallow linguistic analysis includes lemmas, words, part-of-speech tags. TextRunner [43] make use of the dependency parser features of sentences to derive higher quality relation extraction results. Reverb [16] extends TextRunner by enforcing syntactic and lexical constraints on patterns to reduce noisy and incoherent patterns.

Statistical learning based IE. StatSnowball [45] proposes a statistical extraction framework which uses the discriminative Markov logic networks (MLNs) and softens hard rules by learning their weights in a maximum likelihood estimate sense. DeepDive [31] adopts sparse logistic regression ($l1$ regularized) classifiers to train statistical relation-extraction models using both lexical (e.g., word sequences) and syntactic (e.g., dependency paths) features.

3.2 Knowledge Base Construction Approaches

Traditionally, KBs have been constructed manually, typically by experts from a domain. These KBs serve to support decision-making process of humans or machines. However using human experts to construct KBs is not a scalable approach which can be employed to create KBs that do not grow or update frequently. Recently, with the development of semantic web and advanced information extraction technique, a growing number of large scale knowledge bases (KBs), such as YAGO [38], DBpedia [6], Freebase [7], WordNet [30], KnowItALL [15], DeepDive [31], Probase [42], have been constructed. Besides, there are also some commercial projects by Microsoft, Google, Facebook, Walmart and others. These knowledge repositories store millions of facts about the world, such as information about people, places and things (referred as entities).

The automatic knowledge construction can be clustered into 4 main groups:

- Approaches built on Wikipedia infoboxes and other structured data sources, such as YAGO [38], DBpedia [6] and Freebase [7]. The advantage of such approaches is that it can derive a high quality knowledge facts as referring to high quality, structured data sources.

- Approaches aims at constructing taxonomies, which focus on the "isA" relationships and construct the "isA" hierarchies, such as Probase [42]. This is opposed to general KB constructions, as general KB has multiple types of predicates.

- Approaches making use of information extraction (IE) techniques to extract knowledge from schema-less data sources (i.e. web pages, books, etc.), such as Reverb [16], TextRunner [43].

- Approaches aims at knowledge extraction from open sources, which can be applied to the entire web, but use a fixed ontology/schema, such as DeepDive [31].

The summarization and comparison of existing popular knowledge bases are illustrated in Table 1.

4 Crowdsourced Knowledge Acquisition

With increasing usage of the World Wide Web in the recent years, there has been tremendous development in internet-based platforms for collaborative sharing of knowledge. Wikipedia, a collaboratively-edited online encyclopedia, is one of the largest and most popular efforts in this direction. As of date, it contains over 30 million articles in 287 languages (over 4.3 million in English) written, structured, and edited by volunteers on the Web. Its high quality, freely available and dynamically-updated content is widely used for research purposes. Wikipedia is a free and open knowledge base that can be read and edited by both humans and machines. There are "Wikipedians" from all over the world that contributes to edit knowledge. With respect to knowledge acquisition, crowdsourcing techniques also have enjoyed successes. The crowdsourcing platforms offers service to crowdsourced tasks that require human intelligence to complete. In this section, we first introduce existing work about knowledge base construction using the collective wisdom of crowd workers. Furthermore, as there are other kinds of knowledge (called tacit knowledge), which is subjective and experiential, this kind of knowledge resides in human mind and can not be acquired by machine-based techniques and text information inputs. The crowd as a natural source of knowledge can be mine to obtain such knowledge. We will review state-of-the-art approaches for individual knowledge acquisition and commonsense knowledge acquisition.

4.1 Crowdsourced Knowledge Base Construction

Crowdsourcing have attracted much attention in knowledge extraction and taxonomy construction since human power can easily tackle semantic relation recognition and extraction tasks, which are difficult for machines. Some works have been tried to explore the power of crowd for knowledge base construction knowledge acquisition [13, 19, 21].

There are knowledge bases, both of general knowledge and common-sense knowledge, built manually through domain experts or community members. WordNet [30] is a lexical database for the English language which is handcrafted by knowledge engineers. It groups English words into sets of synonyms, provides short definitions and usage examples and

records a number of relations among these synonym sets or their members. Freebase [7] is a collaborative knowledge base consisting of data composed mainly by its community members. Cyc is an artificial intelligence project that attempts to assemble a comprehensive ontology and knowledge base of everyday common sense knowledge, with the goal of enabling AI applications to perform human-like reasoning. The Cyc [27] contains over one million human-defined assertions, rules or common sense ideas, such as "Every tree is a plant" and "Plants die eventually".

There are some works target at taxonomy construction via crowdsourcing. A Taxonomy is a tree structure, denoted as $T = (N, R)$. Each node $e_i \in N$represents an entity (a concept or an instance); each directed edge $r_{ij} \in R$ represents a hypernym/hyponym relation, i.e., $r_{ij} = (e_i, e_j)$ means that e_i is the hypernym of e_j. Cascade [13] proposes an automatic workflow that creates a taxonomy from the collective efforts of crowd workers. Cascade takes a set of items to be categorized and a descriptive phrase identifying these items and generate the taxonomy consisting of labeled categories. The proposed algorithm takes every item and solicits multiple suggested categories for it from different workers. A new set of workers then votes on the best suggested category for each item. Cascade then asks workers to consider every item with all of these "best" categories and judge relevance. Cascade next uses this data to eliminate duplicate and empty categories and to nest related categories, creating a hierarchy. Authors [9] presents DELUGE, an improved workflow that produces taxonomies with comparable quality using significantly less crowd labor. DELUGE is a refinement of the CASCADE algorithm with novel approaches to the subproblems of label elucidation and multi-label classification. For the former, they introduce a combinatorial model that allows us to calculate the relative cost of stopping the label generation phase early. For the problem of classifying items with a fixed set of labels, they present four models: lossless, one-away, a simple probabilistic model, and the MLNB model of label co-occurrence. The latter two models support a greedy control strategy that chooses the most informative label to ask a human to evaluate, within a constant factor of the optimal next label. Authors [19] define a required input from humans in the form of explicit structural, e.g., supertype-subtype relationships between concepts. Through the annotation and voting inputs from

the crowds, they define the principles upon which crowdsourced taxonomy construction algorithms should be based and propose efficient algorithms to construct the taxonomy. Work [20] focus on the domain knowledge base construction via crowdsourcing. Authors aim at build a knowledge base of scientists by crowdsourcing the affiliations of an automatically extracted list of researchers. They extract the authors from a subset of the 2011 DBLP dataset (a database of computer scientists). For each author they construct a multiple choice question of the form X works at Y where X is the author and Y includes three institutions. Users may also submit a free text answer. The results show that find that responses received through Mechanical Turk are more accurate than those received through experts.

4.2 Individual Knowledge Acquisition

In contrast to general knowledge, such as an ontology, knowledge base or information in a database, individual knowledge captures habits and preferences of individuals. As illustrated in the following examples:

- Ann, a vacationer, is interested in finding child-friendly activities at an attraction in NYC, and a good restaurant nearby. For each such combination, she is also interested in useful related advice (e.g., what to wear, whether to walk or rent a bike, etc.). In this case, the *general knowledge* includes: NYC attractions, restaurants, the proximity between them, names of activities; whereas the *individual knowledge* is: what people like doing best with children, what restaurants they like and frequently visit, what they take with them when they go, etc.

- A dietitian wishes to study the culinary preferences in a particular population, focusing on food dishes that are rich in fiber. In this example, the *general knowledge* includes: food dishes that are rich in fiber; whereas the *individual knowledge* is: the culinary preferences of individuals in the population.

- A researcher wishes to study the social habits in communities where the longevity is in the top percentile. *General knowledge*: statistics about the longevity in different communities; and the *individual knowledge* includes the social habits of individuals.

General Knowledge typically refers to the general truth, objective data which can be found in a knowledge base or an ontology; **Individual Knowledge** is related to the habits and opinions of an individual which can be mined from people and crowd. Motivated by the two kinds knowledge and queries with the mixture requirements, authors [4] propose a framework to manage general and individual knowledge in Crowd Mining Applications.

A diagram of the general architecture of a crowd mining framework is given in Figure 4. The Query Engine is responsible for computing an efficient query plan and managing its execution; given a task, it will be sent to multiple workers, the Crowd Task Manager will perform: (i). answer aggregation to overcome crowd errors; (ii). uses a significance function to compute whether sufficient crowd input has been collected, and whether its result is significant, i.e., belongs in the query output; (iii). prioritizes submitting the unfinished tasks to workers based on the estimated effect of more task results on the overall utility. The Inference and Summarization component processes the raw results obtained from the crowd, along with data from the Knowledge Base and User Profile. The Crowd Selection part searches the repository for crowd workers that are suitable to perform a certain task. The selection of workers may be done (i) according to explicit preferences given by the query issuer, (ii) by a similarity between the query issuer and the crowd worker (in tasks which involve recommendations), or (iii) by worker properties related to the task, such as the expertise area, quality and availability of the workers. Users should be able to formulate their information request in a natural language (NL), and this would be automatically translated to the target query language. They also illustrates how the architecture described can used in the OASSIS crowd mining system.

For the crowd mining part, they will ask the crowd questions through single choice HITs, e.g. "How often do you play basketball at Central Park?", given choices "never", "once a week", "once a year", etc. In order to minimize the number of HITs asked, they adopt the framework of crowd mining [5] designed for choosing the best questions to ask the crowd and

mining significant patterns from the answers, making use of the partial order over fact-sets.

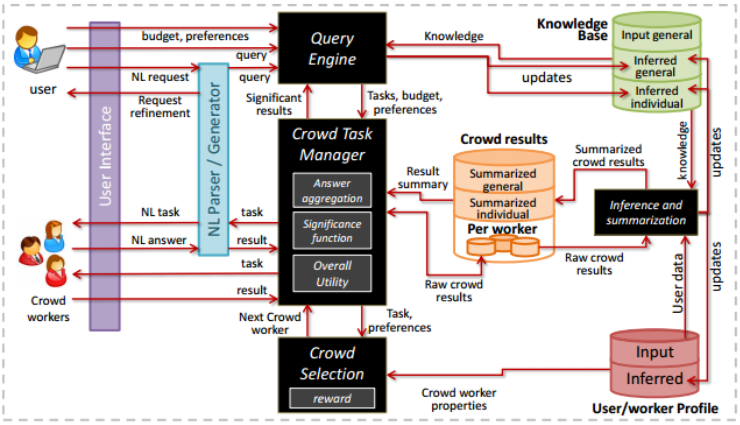


Figure 4: Crowd Mining Architecture

4.3 Commonsense Knowledge Acquisition

In artificial intelligence research, commonsense knowledge is the collection of facts and information that an ordinary person is expected to know. For example, "A lemon is sour", "To open a door", "you must usually first turn the doorknob", "If you forget someone's birthday, they may be unhappy with you". Commonsense knowledge, thus defined, spans a huge portion of human experience, encompassing knowledge about the spatial, physical, social, temporal, and psychological aspects of typical everyday life. Because it is assumed that every person possesses commonsense, such knowledge is typically omitted from social communications, such as text. A full understanding of any text then, requires a surprising amount of commonsense, which currently only people possess.

The commonsense knowledge problem is the ongoing project in the field of knowledge representation (a sub-field of artificial intelligence) to create a commonsense knowledge base: a database containing all the general knowledge that most people possess, represented in a way that it is available to artificial intelligence programs that use natural language or make

inferences about the ordinary world. An early representative example are common sense knowledge bases, with the Open Mind Common Sense (OMCS) project [35], being a prototypical example. OMCS is a knowledge acquisition system designed to acquire commonsense knowledge from the general public over the web, which enabled the construction of a 450,000 assertion commonsense knowledge base. The system acquires facts, descriptions, and stories by allowing participants to construct and fill in natural language templates. It employs word-sense disambiguation and methods of clarifying entered knowledge, analogical inference to provide feedback, and allows participants to validate knowledge and in turn each other. ConceptNet [28] is a large semantic graph of commonsense concepts, related through interlingual and free text relations, These relations relate concepts by their lexical definitions, and through the commonsense associations that ordinary people make. ConceptNet is generated automatically from the English sentences of the Open Mind Common Sense (OMCS) corpus. By applying natural language processing and extraction rules to the semistructured OMCS sentences, 300 000 concepts and 1.6 million binary-relational assertions are extracted to form ConceptNets semantic network knowledge base. Given the concept pencil for example, ConceptNet includes knowledge about the properties that define it (e.g. IsA(pencil, writing instrument)), as well as incidental facts about this concept (e.g. AtLocation(pencil, school), UsedFor(pencil, writing)). SocialExplain [23] aims at acquiring contextual commonsense knowledge from explanations of social content. Incorporating contextual commonsense knowledge into applications can help interpret the observed data for improved reasoning. For example, to answer question "I see a Giants fan and a football helmet. Where am I?", it is necessary to acquire contextual knowledge about the sports domain. The target is to harvest the associations between observations and knowledge used by readers in contextualizing the stream of content on social media. For example, given tweets "Going straight from the lab to The Garden tonight!" followed by "What a great comeback, Celtics rock!", one may postulate that the tweets are from "a basketball-loving college student living in Boston" by reasoning with observations, e.g. "lab", "Garden," and "Celtics", and commonsense knowledge such as "college students work in the lab," "The Celtics are an NBA team," and "The Garden is the home arena for the Boston Celtics." The acquisition process is

[htb] **Figure 6. Question answering interface.**

Figure 5: The Rapport Game

broken into two cognitively simple tasks for human contributors (AMT workers): to identify contextual clues from the given social content, and to explain the content with the clues. [24] introduces the problem of resource-bounded crowdsourcing of commonsense knowledge and proposed an approach to finding a productive question set automatically. By leveraging a guiding knowledge base containing the target domain knowledge, the algorithm generate new questions using a weak form of inference. The candidate questions are sorted according to their expected contributions to the knowledge base and thus being able to improve the acquisition efficiency and concept coverage through optimizing the question selection.

4.4 Games for Knowledge Acquisition

In parallel to "brute-force" crowdsourcing efforts, which relies on volunteers or crowdsourcing platform workers, games with a purpose, have shown that knowledge can be acquired, as a "side-affect", when humans play carefully designed games, termed as Games With A Purpose (GWAP). Human computing games or GWAPs transform or mask computationally challenging tasks into interesting games. There are over millions of game players worldwide, people play games for different reasons, e.g., to relax, to be entertained, for the need of competition and to be thrilled. Additionally, they want to be challenged, both on a mental

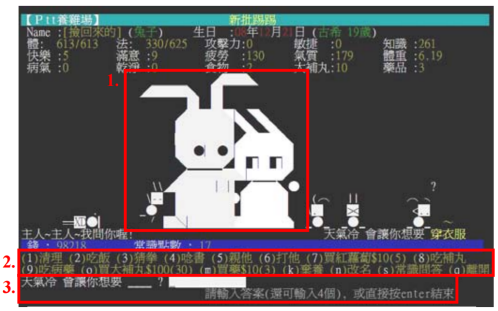


Figure 6: The PTT Game

and on a skill-based level. Such army of gamers could be exploited for performing tasks that are relatively easy to complete by humans, but computationally rather infeasible to solve. The idea is to integrate such tasks as goal of games by producing a win-win situation where people have fun playing games while actually doing something useful. The nature of these games, in fact, focuses on exploiting player inputs to both create meaningful data and provide a funnier game experience The outcome of these games help gather data, which is otherwise hard to acquire (for example, by applying computer vision techniques). Such data is valuable and can directly serve for various purposes (e.g., image search) or indirectly as training data for machine learning approaches.

Some human computing games engage players in knowledge acquisition and knowledge base construction. OntoPronto [36], for example, is a quiz game for vocabulary building that attempts to build a huge domain ontology from Wikipedia articles. This is achieved by mapping random articles to the most specific class of the Proton ontology using the subClassOf relationship. Virtual Pet Game [25] aims to construct a semantic network that encodes common sense knowledge. The game is built on top of PTT, a popular Chinese bulletin board system that is accessible through a terminal interface. Each player owns a pet, which they should take care of by asking and answering questions. The pet in this game is just a substitute for other players, who receive such questions and answers, and have to respond or validate them. As shown in Figure 6, there are Picture with Pets (labeled 1), Choices of player actions (labeled 2) and commonsense question input field.

Rapport Game [25] similarly to Virtual Pet Game, exploits player labor for constructing a semantic network that encodes common sense knowledge, as illustrated in Figure 5. Authors [21] develop a human computing engine that generates game questions where players choose or fill in missing relations for subject-relation- object triples.

Table 1: KB Summarization

KB Name	Information Source	Construction	#Facts	#Instance	#Class	#Relations
YAGO	Wikipedia + WordNet	Automatic	26,120,106	2,747,873	292,898	77
DBpedia	Wikipedia	Automatic	26,770,236	2,526,321	327	1,298
Freebase	Wikipedia + Open Source	Community	2,997 million	40 million	1,450	-
Knowledge Vault	Web Content	Automatic	271 million	45 million	1100	4469
Probase	Web Pages	Automatic	20,757,545	-	2,653,872	-
DeepDive	Web Content	Automatic	7 million	2.7 million	4	34

5 Hybrid system for Knowledge Acquisition

Although advanced techniques have been presented for knowledge acquisition and KB construction, there are fundamental limitations of automatic knowledge acquisition. The automatic methods can yield noisy or semantically meaningless knowledge facts due to the various and complex format, noisy data and semantic ambiguity of the input. Crowdsourcing is a natural alternative to overcome the fundamental limitations of automatic knowledge acquisition. It can help with tasks where fully automatic solutions are deemed inadequate. However, human alone can not carry the whole burden of large scale knowledge acquisition due to the limitation of time, incentive and monetary cost. Motivated by the aforementioned challenges, some hybrid systems which combines the power of machine-based technique and crowdsourcing, have been proposed for better knowledge acquisition and KB construction.

[21] presents a novel system architecture, called Higgins, which shows how to effectively integrate an IE engine and a HC engine. Crowdsourcing is incorporated for relation extraction in the general domain. The IE engine harvests fact candidates from free text by compiling large entity and relation lexicons and using statistics and semantic resources for relational phrases. From the large pool of generated fact candidates, the HC engine uses ranking based on statistical language models to provide most likely candidates for human judgments, casting them either in the form of crowdsourced HITs or questions of a HC game. The architecture is shown in Figure 7.

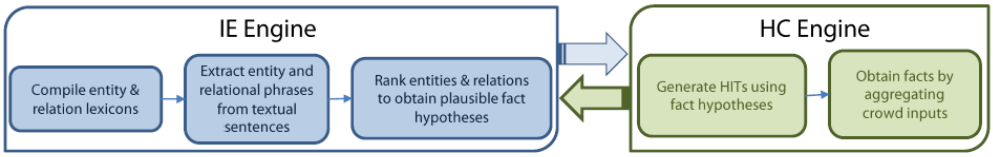


Figure 7: System Architecture of Higgins

In the IE engine, they employ automatic IE on Web corpora, in order to derive candidates for entity-relation-entity triples, with an open set of potential relations. A suite of techniques from computational linguistics, including dependency parsing (with the Stanford

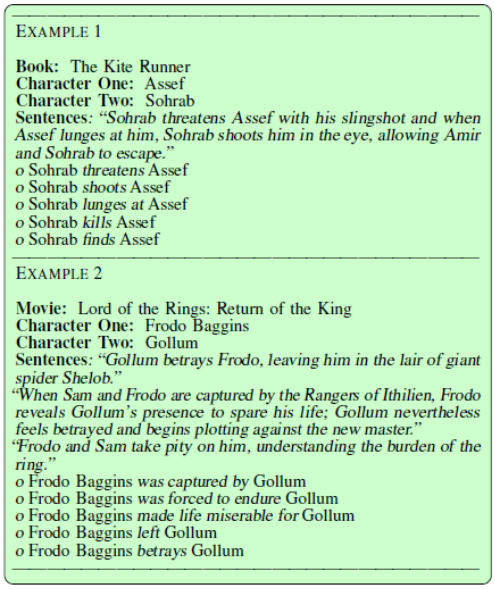


Figure 8: Sample HITs

Parser) and pronoun resolution is adopted. The resulting triples are usually of mixed quality, necessitating the second stage. In the HC engine, the sets of candidates from the IE engine and their underlying patterns are then used to generate HITs in game form. Abstractly, each HIT presents the user with a knowledge quad of the form (c; e_1; r; e_2) where e_1 and e_2 are entities, r is a relation, and c is a cue or textual context. One or more of the components c, e_1, r, and e_2 can be empty slots (variables) to be filled by the user; they may present a multiple-choice list to the user to pick the missing value. The quads are presented in the form of questions, with relevant candidate answers and additional free-text fields for entering further values, illustrated in Figure 8.

In a game, the system could simply ask the user to fill in the missing value for a relationship, by providing a free-text form field. However, experience shows that it is easier to engage a broad set of users with multiple-choice questions. Therefore, they generate a small number of candidate answers (usually 5) and additionally offer a free-text field for entering other relationships. For the game effectiveness, it is important that the offered candidate answers i) are reasonable, that is, exclude nonsense relations for the given context entity

(e.g., exclude is the grandmother of for a question about James Bond and Le Chiffre), ii) include a good answer, if known from the original sentence that generated the HIT, and iii) are sufficiently diverse so that users see actual choices, as opposed to proposing only candidates that are so close that only a very sophisticated user could distinguish them (e.g., orders to kill vs. hires someone to assassinate - which would be considered the same by most players).

Authors [29] present a hybrid framework, which combines the power of machine-based approaches and human computation (the crowd) to construct a more complete and accurate taxonomy, shown in Figure 9. Specifically, the framework consists of two steps: first construct a complete but noisy taxonomy automatically through information extraction technique, then crowd is introduced to adjust the entity positions in the constructed taxonomy to improve the accuracy. However, the adjustment is challenging as the budget (money) for asking the crowd is often limited. To assist the crowd and reduce the burden of adjustment, entities should be selected judiciously to adjust and give candidate positions for each selected entity. Each time, a decision should be made and pick the most beneficial entities to ask and adjust. To evaluate the benefit, they model the utility function, considering both the entity position uncertainty and position amendment benefit; also, each adjustment task is associated with a cost, which is proportional to the number of human intelligence tasks (HITs) needed for the adjustment operation. They formulate the problem of finding the optimal adjustment with limited budget, and then propose an exact algorithm and a more efficient approximation algorithm to solve the problem.

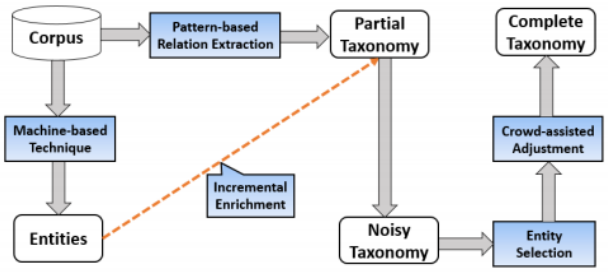


Figure 9: Hybrid Machine-Human Workflow

The work of commonsense knowledge acquisition [24] also investigates the problem of "how to crowdsource effectively within the resource limitation". The resource limitation includes the number of questions or the time limit. They propose a guided crowdsourcing mechanism which identifies the most productive questions and sort the candidate questions according to the expected contributions to the KB. The productiveness is computed through a weak form of inference, i.e. similarity coverage under incomplete information according to a guiding knowledge base. A guiding KB is introduced to help identify the questions by estimating the answers of a question and their inference results before asking users. The algorithm has three steps: (i). Represent every concept in a KB as a feature vector in vector space. For example, dog is a concept and is an animal is its feature. (ii). For every mapped concept in the two KB, find their similar concepts. (iii) If the overlap of the similar concepts in guiding KB and target KB is low, transfer features in guiding KB to create questions for crowd-sourcing. An illustration is shown in Figure 10.

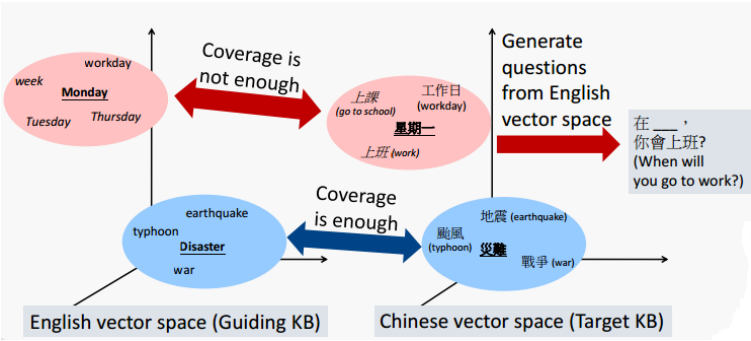


Figure 10: Illustration of Algorithm

6 Summary

In this article, we review the state-of-the-art techniques of crowdsourced knowledge acquisition. Before detailed introduction, we clarify key concepts related to knowledge acquisition and crowdsourcing. For knowledge, there are two types: explicit knowledge (general knowledge) and tacit knowledge, including individual knowledge, commonsense knowledge, etc. Existing approaches for knowledge acquisition can be categorized mainly into three mainstreams: automatic technique, crowdsourcing technique and hybrid techniques which combines the automatic approach and crowdsourcing. We survey the latest development of crowdsourced knowledge acquisition and hybrid system for knowledge acquisition for both explicit knowledge and tacit knowledge.

The increasing popularity of large scale knowledge bases and its driven applications have drawn much attention on the knowledge coverage and accuracy of the KBs. Crowdsourcing as a powerful knowledge source would play more and more important roles in the knowledge acquisition and management. As costs present a scalability bottleneck, hybrid framework is indispensable for scalable, complete and accurate knowledge acquisition. Moreover, there is another kind of knowledge, subjective knowledge, which truth relies on people's dominant opinion, is being neglected. General hybrid frameworks for KA and KB construction and subjective knowledge acquisition are two possible directions forward from current work.

References

[1] Amazon mechanical turk. http://www.mturk.com.

[2] Crowdflower. http://crowdflower.com.

[3] Eugene Agichtein and Luis Gravano. Snowball: Extracting relations from large plain-text collections. In *Proceedings of the fifth ACM conference on Digital libraries*, pages 85–94. ACM, 2000.

[4] Yael Amsterdamer, Susan B. Davidson, Anna Kukliansky, Tova Milo, Slava Novgorodov, and Amit Somech. Managing general and individual knowledge in crowd mining applications. In *CIDR 2015, Seventh Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, January 4-7, 2015, Online Proceedings*, 2015.

[5] Yael Amsterdamer, Yael Grossman, Tova Milo, and Pierre Senellart. Crowd mining. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, pages 241–252. ACM, 2013.

[6] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary G. Ives. Dbpedia: A nucleus for a web of open data. In *ISWC +ASWC*, 2007.

[7] Kurt D. Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *SIGMOD*, 2008.

[8] Jonathan Bragg, Mausam, and Daniel S. Weld. Crowdsourcing multi-label classification for taxonomy creation. In *HCOMP*, 2013.

[9] Jonathan Bragg, Daniel S Weld, et al. Crowdsourcing multi-label classification for taxonomy creation. In *First AAAI conference on human computation and crowdsourcing*, 2013.

[10] Sergey Brin. Extracting patterns and relations from the world wide web. In *International Workshop on The World Wide Web and Databases*, pages 172–183. Springer, 1998.

[11] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka, Jr., and Tom M. Mitchell. Toward an architecture for never-ending language learning. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, AAAI'10, pages 1306–1313. AAAI Press, 2010.

[12] Lydia B. Chilton, Greg Little, Darren Edge, Daniel S. Weld, and James A. Landay. Cascade: crowdsourcing taxonomy creation. In *CHI*, 2013.

[13] Lydia B Chilton, Greg Little, Darren Edge, Daniel S Weld, and James A Landay. Cascade: Crowdsourcing taxonomy creation. In *CHI*, 2013.

[14] Xu Chu, John Morcos, Ihab F. Ilyas, Mourad Ouzzani, Paolo Papotti, Nan Tang, and Yin Ye. KATARA: A data cleaning system powered by knowledge bases and crowdsourcing. In *SIGMOD*, 2015.

[15] Oren Etzioni, Michael J. Cafarella, Doug Downey, Stanley Kok, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. Web-scale information extraction in knowitall: (preliminary results). In *Proceedings of the 13th international conference on World Wide Web, WWW 2004, New York, NY, USA, May 17-20, 2004*, pages 100–110, 2004.

[16] Anthony Fader, Stephen Soderland, and Oren Etzioni. Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545. Association for Computational Linguistics, 2011.

[17] Ju Fan, Meiyu Lu, Beng Chin Ooi, Wang-Chiew Tan, and Meihui Zhang. A hybrid machine-crowdsourcing system for matching web tables. In *ICDE*, 2014.

[18] Marti A Hearst. Automatic acquisition of hyponyms from large text corpora. In *COLING*, 1992.

[19] Dimitris Karampinas and Peter Triantafillou. Crowdsourcing taxonomies. In *ESWC*. 2012.

[20] Ari Kobren, Thomas Logan, Siddarth Sampangi, and Andrew McCallum. Domain specific knowledge base construction via crowdsourcing.

[21] Sarath Kumar Kondreddi, Peter Triantafillou, and Gerhard Weikum. Combining information extraction and human computing for crowdsourced knowledge acquisition. In *ICDE*, 2014.

[22] Sarath Kumar Kondreddi, Peter Triantafillou, and Gerhard Weikum. Combining information extraction and human computing for crowdsourced knowledge acquisition. In *ICDE*, 2014.

[23] Yen-Ling Kuo, J Hsu, and Fuming Shih. Contextual commonsense knowledge acquisition from social content by crowd-sourcing explanations. In *Proceedings of the Fourth AAAI Workshop on Human Computation*, pages 18–24, 2012.

[24] Yen-Ling Kuo and Jane Yung-jen Hsu. Resource-bounded crowd-sourcing of commonsense knowledge. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, volume 22, page 2470, 2011.

[25] Yen-ling Kuo, Jong-Chuan Lee, Kai-yang Chiang, Rex Wang, Edward Shen, Cheng-wei Chan, and Jane Yung-jen Hsu. Community-based game design: experiments on social games for commonsense data collection. In *Proceedings of the acm sigkdd workshop on human computation*, pages 15–22. ACM, 2009.

[26] Matthew Lease and Omar Alonso. Crowdsourcing and human computation, introduction. In *Encyclopedia of Social Network Analysis and Mining*. 2014.

[27] Douglas B Lenat. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38, 1995.

[28] Hugo Liu and Push Singh. Conceptneta practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226, 2004.

[29] Rui Meng, Yongxin Tong, Lei Chen, and Caleb Chen Cao. CrowdTC: Crowdsourced taxonomy construction. In *ICDM*, 2015.

[30] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 1995.

[31] Feng Niu, Ce Zhang, Christopher Ré, and Jude W. Shavlik. Deepdive: Web-scale knowledge-base construction using statistical learning and inference. In *Proceedings of the Second International Workshop on Searching and Integrating New Web Data Sources, Istanbul, Turkey, August 31, 2012*, pages 25–28, 2012.

[32] Alexander J Quinn and Benjamin B Bederson. Human computation: a survey and taxonomy of a growing field. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1403–1412. ACM, 2011.

[33] Alan Ritter, Stephen Soderland, and Oren Etzioni. What is this, anyway: Automatic hypernym discovery. In *AAAI*, 2009.

[34] Feng Shi, Juanzi Li, Jie Tang, Guo Tong Xie, and Hanyu Li. Actively learning ontology matching via user interaction. In *ISWC*, 2009.

[35] Push Singh, Thomas Lin, Erik T Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu. Open mind common sense: Knowledge acquisition from the general public. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pages 1223–1237. Springer, 2002.

[36] Katharina Siorpaes and Martin Hepp. Ontogame: Weaving the semantic web by online games. In *European Semantic Web Conference*, pages 751–766. Springer, 2008.

[37] Rion Snow, Daniel Jurafsky, and Andrew Y Ng. Learning syntactic patterns for automatic hypernym discovery. In *NIPS*, 2004.

[38] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. YAGO: A large ontology from wikipedia and wordnet. *J. Web Sem.*, 2008.

[39] Norases Vesdapunt, Kedar Bellare, and Nilesh N. Dalvi. Crowdsourcing algorithms for entity resolution. *PVLDB*, 2014.

[40] Jiannan Wang, Tim Kraska, Michael J. Franklin, and Jianhua Feng. CrowdER: Crowdsourcing entity resolution. *PVLDB*, 2012.

[41] Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Q Zhu. Probase: A probabilistic taxonomy for text understanding. In *SIGMOD*, 2012.

[42] Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Q. Zhu. Probase: A probabilistic taxonomy for text understanding. In *SIGMOD*, 2012.

[43] Alexander Yates, Michael Cafarella, Michele Banko, Oren Etzioni, Matthew Broadhead, and Stephen Soderland. Textrunner: open information extraction on the web. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 25–26. Association for Computational Linguistics, 2007.

[44] Chen Jason Zhang, Lei Chen, H. V. Jagadish, and Caleb Chen Cao. Reducing uncertainty of schema matching via crowdsourcing. *PVLDB*, 2013.

[45] Jun Zhu, Zaiqing Nie, Xiaojiang Liu, Bo Zhang, and Ji-Rong Wen. Statsnowball: a statistical approach to extracting entity relationships. In *Proceedings of the 18th international conference on World wide web*, pages 101–110. ACM, 2009.