**Mining Knowledge on Relationships between Objects from the Web**

Xinpeng ZHANG†(a), Nonmember, Yasuhi ASANO†, and Masatoshi YOSHIKAWA†, Members

**SUMMARY** How do global warming and agriculture influence each other? It is possible to answer the question by searching knowledge about the relationship between global warming and agriculture. As exemplified by this question, strong demands exist for searching relationships between objects. Mining knowledge about relationships on Wikipedia has been studied. However, it is desired to search more diverse knowledge about relationships on the Web. By utilizing the objects constituting relationships mined from Wikipedia, we propose a new method to search images with surrounding text that include knowledge about relationships on the Web. Experimental results show that our method is effective and applicable in searching knowledge about relationships. We also construct a relationship search system named “Enishi” based on the proposed new method. Enishi supplies a wealth of diverse knowledge including images with surrounding text to help users to understand relationships deeply, by complementarily utilizing knowledge from Wikipedia and the Web.

**key words:** knowledge retrieval, Wikipedia mining, content-based image retrieval, relationships between objects

1. Introduction

What is the relationship between petroleum and Japan? Why has the bankruptcy of Lehman Brothers Holdings Inc. so strongly influenced Japan? How global warming and agriculture mutually affect each other? Answering these questions demands knowledge about relationships between objects, such as countries, products, people, and events. Our real life includes strong needs for querying knowledge about relationships. For example, understanding the relationship between two countries is useful to study history or politics. Discovering the relationship between a country and a product can presumably help in making an investment decision. Surveying the connection between oneself and an unknown person is necessary to find paths to contact the unknown person through some mutual acquaintances. Knowing the existence and the strength of a relationship is a way for understanding a relationship to some extent. To understand a relationship deeply, we need to know how the relationship is established. In this paper, to help users to understand relationships deeply, we address the problem of searching diverse knowledge about relationships on the Web.

In Wikipedia, the knowledge associated with an object is well organized on a single page. Wikipedia contains knowledge of objects in numerous categories such as people, science, geography, politics, and history. The link structure between Wikipedia pages represents the relationships between objects. Therefore, Wikipedia is efficient for searching knowledge about objects and relationships between objects. For example, Zhang et al. [2], [3] proposed methods to extract paths from the link structure of Wikipedia for explaining relationships. However, the knowledge contained in Wikipedia is still much less than that existing on the Web. Especially, multimedia information such as images is lacking in Wikipedia.

It is desired to search more knowledge about relationships, including images on the Web, to complement the knowledge available in Wikipedia. However, the knowledge existing on the Web is not well organized, in general (We consider Wikipedia separately from the Web in this paper). A Web page might contain knowledge related to multiple objects; multiple different objects might be represented by the same word. It is a difficult challenge to search knowledge about relationships between objects on the Web. Recently, several semantic knowledge bases [4], [5] are used for searching semantic relationships between two objects. However, the semantic relationships defined in these knowledge bases, such as “isCalled,” “type,” and “sub-ClassOf,” are far from covering relationships existing in the real world. For example, questions described at the beginning of this section are intended to obtain diverse knowledge about relationships, rather than simple semantic relationships. Searching diverse knowledge about relationships on the Web has not been well studied. Especially, no method exists for searching multimedia information about relationships, including images, on the Web.

In this paper, we propose an evidence-based method, to search sets of an image with surrounding text (hereafter abbreviated as IwST) that include knowledge (i.e., descriptions) about relationships. For example, Fig. 2 presents an IwST searched by our method that includes the knowledge about the relationship between “Global warming” and “Agriculture.” The image presented in Fig. 2 shows concrete figures about how much the agricultural productivity will tail in different areas and countries by 2080s, if carbon emissions continue; while its surrounding text explains why and how the concrete figures are calculated. It is better to read the image and its surrounding text together to gain more knowledge about the relationship rather than reading the image or the text only. Therefore, we search an image with its surrounding text as a set rather than searching images or text.
only. We discuss more about the issue in Sect. 2.2.

Two kinds of relationships exist: “explicit relationships,” and “implicit relationships” [6]. The evidence-based method searches knowledge for implicit relationships. An explicit relationship is a direct relationship of two objects, such as A is a “friend” of B. An implicit relationship is an indirect relationship of two objects established through the intermediary of other objects, such as A is a “friend of a friend” of C through B who is a friend of both A and C. As another example of an implicit relationship, “carbon dioxide” is one of the objects constituting the relationship between “Global warming” and “Agriculture,” corresponding to the fact that global warming and agriculture affect each other through carbon dioxide. We call the intermediary objects forming a relationship “elucidatory objects.” It is possible to obtain elucidatory objects of a relationship on Wikipedia using the method [2], [3] proposed by Zhang et al. Our evidence-based method then utilizes elucidatory objects to search IwSTs for implicit relationships. Concretely, our method estimates that an IwST includes knowledge about an implicit relationship, if many elucidatory objects of the relationship appear in the surrounding text of the image. For example, the IwST presented in Fig. 2 includes the knowledge about the relationship between “Global warming” and “Agriculture.” As shown in Fig. 2, many elucidatory objects of the relationship appear in the surrounding text. We confirm through experiments described in Sect. 4 that the evidence-based method is effective for searching IwSTs that include knowledge about relationships.

We also construct a relationship search system named “Enishi.” Enishi utilizes the knowledge in Wikipedia and that on the Web complementarily to search knowledge about relationships. Generally, relationships existing in the real world are complicated and diverse. To help users to deeply understand relationships between pairs of objects, Enishi supplies a wealth of knowledge about relationships. Given two objects, Enishi (a) mines paths formed by links in Wikipedia that explain the relationship between the two objects; (b) searches IwSTs that include knowledge about the relationship on the Web; and (c) offers a user interface showing results of (a) and (b). Figure 1 depicts the system architecture of Enishi. We use the existing method proposed by Zhang et al. [2], [3] to (a) mine paths on the Wikipedia link structure. For example, Fig. 3 shows 10 paths mined for the relationship between “Global warming” and “Agriculture.” Users can obtain knowledge about the relationship to some extent through understanding the mined paths. The intermediate objects existing in the paths mined for a relationship, are the elucidatory objects of the relationship. We then propose approaches based on the evidence-based method to (b) search IwSTs that include knowledge about a relationship on the Web, by utilizing the elucidatory objects in the paths mined in (a). By reading the IwSTs, users can gain more knowledge about the relationship than reading the paths minded in Wikipedia only. Consequently, Enishi is applicable to support users to deeply understand relationships. To the best of our knowledge, Enishi is the first system for searching diverse knowledge about relationships, especially images. Although there are services on the Web for searching people related to certain people [7]–[9].

The main contributions of this paper are as follows:

---

1Enishi is a Japanese word meaning relationship, in this paper. The system is accessible at http://www.db.soc.i.kyoto-u.ac.jp/enishi/enishi.html.
1. We construct Enishi system for searching diverse knowledge about relationships from both Wikipedia and the Web (Sect. 2).
2. We propose an evidence-based method for searching InSTs that include knowledge about a relationship on the Web. We then propose approaches based on the evidence-based method for Enishi system to search InSTs about relationships (Sect. 3).

We report experimental results in Sect. 4. We review related work in Sect. 5. The method proposed by Zhang et al. [2],[3] for mining relationships in Wikipedia is introduced in Appendix.

2. Enishi Search System

The Enishi (relationship) Search System searches knowledge of the following three types about a relationship.

(1) Disjoint paths formed by links in Wikipedia that are important for explaining the relationship.
(2) Images with surrounding text retrieved on the Web that include knowledge about the overall relationship.
(3) Images with surrounding text retrieved on the Web that include knowledge about each path for the relationship of type (1).

Figure 3 portrays the interface of Enishi, in which we search the relationship between “Global warming” and “Agriculture.” Enishi is implemented based on Wikipedia. It searches the relationship between two objects, each of which is designated by the title of the Wikipedia page representing the object. We discuss the knowledge of each type at below.

2.1 Disjoint Paths Mined from Wikipedia

Two kinds of relationships exist: “explicit relationships” and “implicit relationships [6],” as discussed in Sect. 1. In Wikipedia, an explicit relationship is represented as a link. For example, an explicit relationship between global warming and carbon dioxide might be represented by a link from the page “Global warming” to the page “Carbon dioxide.” A user could understand its meaning by reading the text “greenhouse gases have a mean warming effect...” The major greenhouse gases are... carbon dioxide...” surrounding the anchor text “carbon dioxide” on page “Global warming.” In Wikipedia, multiple links and pages represent an implicit relationship. For example, global warming affects agriculture via carbon dioxide. This fact might be an implicit relationship represented by two links in Wikipedia: one between “Global warming” and “Carbon dioxide” and the other one between “Agriculture” and “Carbon dioxide.”

We define a Wikipedia information network, whose vertices are pages of Wikipedia and whose edges are links between pages. We then mine disjoint paths between two objects on a Wikipedia information network for explaining the relationship between the two objects using the method proposed by Zhang et al. [3]. The paths mined for explaining a relationship satisfy the following two requirements, (1) the paths are useful for understanding the relationship, and (2) the middle objects in the paths play important roles in the relationship. Every path is assigned with a score representing its importance to explain a relationship. Enishi displays the top 10 important paths for a relationship. For example, Fig. 3 shows the top 10 paths for the relationship between “Global warming” and “Agriculture.” A user can understand the meaning of a path easily by tracing the links in the path from left to right. Tracing each link can be done by understanding the meaning of an explicit relationship represented by the link. For each link (u, v), Enishi extracts a snippet surrounding the anchor text of link v on page u for explaining the link. In Enishi, a balloon displaying the snippet for a link is popped up by a mouse roll over on the link. For example, in Fig. 3, if users read the snippets indicated by the number 1, then they can understand the bottom path containing “Carbon dioxide.”

Users prefer not to read the same objects repeatedly in the mined paths, and they might desire to obtain knowledge of various kinds by understanding a few paths. Therefore, Zhang et al. [3] mine disjoint paths to avoid outputting redundant objects in the mined paths. In Wikipedia, a path constituted by edges of different directions might be important for a relationship, such as the bottom path in Fig. 3. Therefore, Zhang et al. [3] mine both paths constituted by edges of any direction. Zhang et al. [3] confirmed through experiments that the method can mine many disjoint paths useful for understanding relationships.

2.2 Images with Surrounding Text for the Overall of a Relationship

It is desired to search more knowledge about a relationship on the Web, especially images lacking in Wikipedia. Enishi searches sets of “image with surrounding text” including knowledge about a relationship on the Web, using methods which will be discussed in Sect. 3. An image itself might make it difficult to infer a relationship. The surrounding text of an image is relevant to the image in most cases, as shown in the text based image retrieval researches [10],[11]. By reading an image with its surrounding, users were able to understand the image, and to gain knowledge described in the surrounding text. For example, Fig. 4 portrays an image in which both George W. Bush and Junichiro Koizumi appear. It is difficult to infer the relationship between the two by watching the image. By reading the surrounding text of the image, we were able to understand that they had dis-

![Image of George W. Bush and Junichiro Koizumi](image_url)
cussed international issues together. On the other hand, the image could be an evidence of the fact described in the surrounding text, and is interesting and helpful for understanding the text. Therefore, we search images with surrounding text together.

In Enishi, a window is popped up by pressing the button “Flip View,” on which users were able to browse images with surrounding text associated with the overall of a relationship, as indicated by the number 2. As another manner for viewing images associated with a relationship, by pressing the button “Tile View,” images are displayed one after the other on the corners, as indicated by the number 3.

2.3 Images with Surrounding Text for Each Path

We extract snippets from Wikipedia to explain every link in a path mined from Wikipedia. A user can understand the meaning of a path to some extent by reading the snippets for the links in the path. Searching more knowledge about a path is useful for deepening understanding of the path. Therefore, Enishi also searches images with surrounding text on the Web particularly for explaining a path. Precisely, for a path of a relationship between objects \( s \) and \( t \), Enishi searches \( IwSTs \) including knowledge about the relationships between objects in each link of the path, or knowledge about how important the intermediate objects in the path are to the relationship between \( s \) and \( t \). For example, the \( IwST \) of ID 2 presented in Fig. 7 includes knowledge about current sea level rise as a cited proofs of global warming. The \( IwST \) is useful for deeply understanding the path containing “Current sea level rise” of the relationship between “Global warming” and “Agriculture.”

In Enishi, a window is popped up by clicking a path. Thereby, users were able to browse images associated with the path on the windows. For example, in Fig. 3, the window indicated by number 4 displays an image for the top path containing “Energy development.”

3. Searching \( IwSTs \) including Knowledge about Relationships on the Web

In this section, we first propose an evidence-based method for searching \( IwSTs \) associated with a relationship. Based on the method, we describe a Top-Down approach for searching \( IwSTs \) that explain the overall of a relationship, and a Bottom-Up approach for searching \( IwSTs \) that explain each path of a relationship.

3.1 Evidence-Based Method

We propose an evidence-based method for searching \( IwSTs \) associated with a relationship. The surrounding text of an image is considered as relevant information describing the image, in several image retrieval system [10], [11]. Similarly, if the surrounding text of an image includes descriptions about the relationship, our evidence-based method regards the image as relevant to the relationship. Our method then searches \( IwSTs \) for a relationship by analyzing the surrounding text of images as discussed in the following.

The method proposed by Zhang et al. [3] mines the top-\( k \) paths that are important for explaining a relationship between objects \( s \) and \( t \) in Wikipedia. The objects in the top-\( k \) paths, except \( s \) and \( t \), are elucidatory objects. Zhang et al. [3] ascertained through experiments that many elucidatory objects in the top-\( k \) paths mined for a relationship, play important roles in constituting the relationship. For example, as illustrated in Fig. 3, “Carbon dioxide” is one of the elucidatory objects playing important roles in the relationship between “Global warming” and “Agriculture.” We assign a weight \( 0 < w(o) < 1 \) to every elucidatory object \( o \) in a path \( p \), using the value of the decomposed flow on path \( p \), which represents the importance of \( p \). A high weight signifies that the elucidatory object plays an important role in constituting a relationship. We observe that elucidatory objects are useful evidences for judging whether a text includes knowledge about a relationship. As an example, let look at the image and its surrounding text presented in Fig. 2 again, which explain how global warming affects agriculture through temperature rise and carbon emissions. Several elucidatory objects of the relationship between “Global warming” and “Agriculture” appear many times in the surrounding text, such as “Temperature,” “Carbon dioxide,” “Greenhouse gas,” and “Crop yield.” Our evidence-based method infers that an \( IwST \) includes knowledge about the relationship between \( s \) and \( t \), if the surrounding text of the image contains two objects \( s \) and \( t \), and many elucidatory objects of the relationship.

It is inefficient to search \( IwSTs \) containing elucidatory objects of a relationship on the whole Web. Instead, we search \( IwSTs \) from the result obtained using a keyword image search engine. Keyword image search engines, such as Google image search engine\(^1\), offer us an easy way for searching images related to two objects. For example, we can search images related to “petroleum” and “Japan” using a query “petroleum Japan,” or “petroleum and Japan.” Some result images and their surrounding text generated using these queries include knowledge about the relationship between “petroleum” and “Japan.” However, some images include no knowledge about the relationship, such as images of products using oil from shopping sites, images of books talking about petroleum, or images even almost unrelated to “petroleum” or “Japan.” The engine returns images whose pages containing words in the queries, but it does not search images associated with a relationship. Our evidence-based method finds \( IwSTs \) containing elucidatory objects of a relationship from the result images.

We describe the method \( \text{EBM}(s, t, k, m, n) \) for searching \( IwSTs \) about the relationship between \( s \) and \( t \).

\textbf{Input:} objects \( s \) and \( t \), integer parameters \( k, m, \) and \( n \).

(1) Mine the top-\( k \) paths that are important for the relationship between \( s \) and \( t \) using the method discussed in Sect. A.2.

\(^1\)http://images.google.com/
(2) Obtain a set $O$ of elucidatory objects in the top-$k$ paths.
(3) Search the top-$m$ images, say $m = 300$, using a key-
word image search engine with query “s t.”
(4) Extract the surrounding text of each image. Let $I$ be
the set of the top-$m$ images with surrounding text.
(5) Remove $lwST$s whose surrounding text contains no $s$
or $t$ from $I$.
(6) Compute a score $s(i)$ for every $i \in I$ to
\[
    s(i) = \sum_{o \in I, o'} w(o) \cdot \log_e(e + f(o)), \tag{1}
\]
where $O' \subseteq O$ is the set of elucidatory objects appear-
ing in the surrounding text of $i$, and $f(o)$ is the appear-
ance frequency of $o$ in $i$. The weight $w(o)$ of $o$ is equal
to the value of the decomposed flow on the path where $o$
appears. The weight $w(s)$ and $w(t)$ is set to the maxi-
num weight of all objects in $O$.
(7) Output: the top-$n$ $lwST$s $i \in I$ having high scores.

We implement a DOM-Tree based method which is
much the same as the one proposed by Fauzi et al. [12], to
extract surrounding text for images at step (4). The method
works well in most cases.

An object might be represented by multiple alternative
words on the Web. For example, “USA,” “U.S.A.,” “U.S.,”
“America,” and “United States of America,” all represent
the “United States.” Our method counts the appearance fre-
cquency of an object including the instances of its alternative
words at step (6). We obtain alternative words represent-
ing an object using redirect pages in Wikipedia. For exam-
ple, searching “USA” will redirect you to the page “United
States” in Wikipedia.

A text in which many elucidatory objects appear a few
times tends to include more diverse knowledge about a re-
lationship, than a text in which few elucidatory objects ap-
pear many times. We take the logarithm of the appear-
fance frequency in Eq. (1) to alleviate the influence of the high
appearance frequency of a single elucidatory object.

3.2 Top-Down Approach

As discussed in Sect. 2, the Enishi system searches both
$lwST$s explaining the overall of a relationship and $lwST$s
explaining each path that is important for a relationship
mined from Wikipedia. In this section, using the evidence-
based method, we propose a Top-Down approach that first
searches a set of $lwST$s for the overall relationship, and
then select $lwST$s from those for the overall relationship to
explain each path of the relationship. We present the Top-
Down approach as follows.

Input: objects $s$ and $t$, integer parameters $k$, $m$, $n$, and

(1) Search the top-$n$ $lwST$s associated with the relation-
ship $r$ between $s$ and $t$ using the evidence-based method
$EBM(s, t, k, m, n)$. Let $TI$ be the set of the top-$n$
$lwST$s.
(2) Output: $TI$ as the $lwST$s for explaining the overall of
relationship $r$.
(3) To search $lwST$s for each path $p$ that is important for
relationship $r$, do step (4) and (5).
(4) Compute a score $s(i)$ for each $i \in TI$ toward path $p$
using the following equations. Let $O'_p \subseteq O$ be the set
of elucidatory objects which exist in path $p$ and appear
in the surrounding text of $i$.
\[
    s'(i) = \sum_{o \in O'_p} w(o) \cdot \log_e(e + f(o))
\]
\[
    s(i) = \begin{cases} 
    0 & (s'(i) = 0) \\
    s'(i) + \sum_{o \in I} w(o) \cdot \log_e(e + f(o)) & (s'(i) \neq 0)
    \end{cases}
\]
(5) Output: the top-$l$ $i \in TI$ having high scores $s(i)$ as the
$lwST$s explaining path $p$.

The Top-Down approach first searches a set $TI$ of
$lwST$s for the overall relationship using the evidence-based
method. The approach then designates an $lwST$ in $TI$ for
a path if the surrounding text of the $lwST$ contains eluci-
datory objects in the path. If the surrounding text of any $lwST$
in $TI$ contains no elucidatory object in a path, then the Top-
Down approach can not search $lwST$s for the path.

3.3 Bottom-Up Approach

We present the Bottom-Up approach in the following. It first
searches $lwST$s for each path, then constructs $lwST$s for the
overall relationship from the $lwST$s for each path.

Input: objects $s$ and $t$, integer parameters $k$, $m$, $n$, and

(1) For each top-$k$ path $p = (o_0 = s, o_1, \ldots, o_n = t)$ that
is important for the relationship $r$ between $s$ and $t$, do
step (2) and (3).
(2) For each edge $(o_i, o_{i+1})$ in $p$, search the top-$n$ $lwST$s
associated with the relationship between $o_i$ and $o_{i+1}$ using
the evidence-based method $EBM(o_i, o_{i+1}, k, m, n)$.
(3) Output: the top-$l$ $lwST$s having high scores among
the $lwST$s for every edge in $p$, as the $lwST$s explaining
path $p$.
(4) Output: the top-$n$ $lwST$s having high scores among
the $lwST$s for every top-$k$ path, as the $lwST$s explain-
ing the overall of relationship $r$.

The Bottom-Up approach first searches $lwST$s for each
of all the edges in a path, as the $lwST$s explaining the path.
The approach then selects those having high scores from the
$lwST$s explaining each of all paths, as the $lwST$s explaining
the overall of a relationship. We compare the Top-Down
method and the Bottom-Up method through experiments in
Sect. 4.

4. Experiments

In this section, we conduct experiments to evaluate the ap-
proaches discussed in Sect. 3 for searching $lwST$s associ-
ated with a relationship on the Web.
4.1 Dataset

We select 40 relationships between two objects of the following four types for evaluation: (1) global warming and an industry, (2) petroleum and a country, (3) a politician and a country, and (4) two politicians. From a English Wikipedia dataset (2010/03/12 snapshot) containing 11,930,677 pages and 140,937,586 links between two pages, we mined the top-150 disjoint paths important for every relationship using the method proposed by Zhang et al. [3]. To search $IuSTs$ for the overall and each of the top-5 paths of the 40 relationships, we totally gathered 400,000 images with their Web pages from the Web.

4.2 Evaluation of Searching Knowledge on the Web

We consider the following questions to evaluate the Top-Down and the Bottom-Up approaches.

Q1. Do the $IuSTs$ searched by each of the two approaches include knowledge about the overall of a relationship or knowledge about each path of a relationship? Which approach is better?

Q2. Do the $IuSTs$ searched on the Web include new knowledge unavailable in Wikipedia?

As discussed in Sect. 2, we search an image with its surrounding text as a set that includes knowledge about a relationship. Therefore, we do not evaluate obtained images separately from their surrounding text.

4.2.1 Human Subjects

We answer questions (Q1) using evaluations by human subjects. We randomly select 8 from the 40 relationships (two from each of the four types) explained above, which are listed in Table 1. For each relationship, we output the top-10 $IuSTs$ for the overall relationship and the top-5 $IuSTs$ for each of the top-5 paths of the relationship, respectively using the Top-Down approach and the Bottom-Up approach. We use the objects in the top-50 paths as elucidatory objects for both approaches. The evidence-based method outputs $IuSTs$ associated with a relationship between $s$ and $t$ from the top-300 images searched using Google image search engine using a query “s t.” We regard the top-10 images searched using Google for a relationship as the $IuSTs$ explaining the overall relationship outputted by Google; we compare the $IuSTs$ with those searched using our approaches. Totally, we obtained 417 $IuSTs$ for the eight relationships (different approaches output some identical $IuSTs$).

We then ask 10 testers to evaluate the 417 $IuSTs$. To each $IuST$ for the overall relationship, every tester assigns an integer score of 0, 1, or 2 representing whether the $IuST$ includes descriptions (i.e. knowledge) about the relationship. To each $IuST$ for a path of a relationship, every tester assigns an integer score of 0, 1, or 2 representing whether the $IuST$ includes descriptions (i.e. knowledge) about any relationship represented by each edge in the paths, or descriptions (i.e. knowledge) about the importance of the elucidatory objects in the path to the relationship. A higher score represents an $IuST$ includes more knowledge about the overall or a path of a relationship. Score 0 means that an $IuST$ includes no knowledge about a relationship. Each tester assigns scores independently of the others by reading every image with its surrounding text. We then compute the average of the scores given by the 10 testers of every $IuST$ searched for each relationship.

Figure 5 presents the average scores of the $IuSTs$ for each overall relationship obtained using each approach. A bar is combined by the lower dark-colored part and the upper light-colored part. Let $S(i)$ be the sum of scores $i \in 0, 1, 2$ assigned by the testers to the $IuSTs$ outputted by a approach for a relationship. The dark-colored part and the light-colored part indicate the proportion $p(2) = \frac{S(2)}{S(2)+S(1)}$ and $p(1) = \frac{S(1)}{S(2)+S(1)}$, respectively. The Top-Down approach produced the highest average for all relationships. As shown by the dark-colored parts of the bars, the $IuSTs$ outputted by the Top-Down approach also received more scores 2 (on average for the 8 relationships $p(2) = 0.69$) than other approaches. By contrast, 75.5% of the scores assigned to the $IuSTs$ for all the 8 relationships searched using Google are 0 or 1. For example, many top-10 images searched using Google for the relationship between George W. Bush and Junichiro Koizumi are assigned score 0. The surrounding text of these images includes almost no description about the relationship, although the two politicians appear in the images. As another example, for the relationship between petroleum and Japan, the surrounding text of many top-10 images searched using Google include knowledge related only to petroleum or Japan, or include no knowledge about

Table 1  Selected relationships for human subjects.

| ID | Relationship               |
|----|---------------------------|
| 1  | Global warming - Agriculture |
| 2  | Global warming - Energy industry |
| 3  | Petroleum - Saudi Arabia |
| 4  | Petroleum - Japan         |
| 5  | Vladimir Putin - Ukraine |
| 6  | Hu Jintao - North Korea  |
| 7  | Jacques Chirac - Wen Jiabao |
| 8  | George W. Bush - Junichiro Koizumi |
the relationship at all. The performance of the Bottom-Up approach was inferior. The Bottom-up approach searches \textit{IuST}s for an overall relationship from the \textit{IuST}s explaining each path of the relationship. Testers consider that most \textit{IuST}s explaining a path include little knowledge about the overall relationship. For example, the \textit{IuST} of ID 2 presented in Fig. 7 explains the path containing “Current sea level rise” of the relationship between “Global warming” and “Agriculture.” However, the \textit{IuST} includes little knowledge about the overall of the relationship between “Global warming” and “Agriculture.” We conclude that the Top-Down approach is sufficient for searching \textit{IuST}s for the overall relationship.

Figure 6 presents the average scores of the \textit{IuST}s of the top 5 paths of each relationship. Similarly to Fig. 5, the dark-colored part indicates the proportion \( \frac{S(2)}{S(2)+S(1)} \). The Bottom-Up approach produced much higher average than the Top-Down approach did for all relationships except the relationship of ID 6 between Hu Jintao and North Korea. The Bottom-Up approach also gained more scores 2 than the Top-Down approach. The Top-Down approach searches \textit{IuST}s for a path of a relationship among \textit{IuST}s explaining the overall relationship, however, experiment results revealed that many of these \textit{IuST}s includes little or no knowledge strongly related to a single path. On the other hand, the Bottom-Up approach can search \textit{IuST}s for a path that include knowledge about any relationship represented by each edge in the path. For example, Fig. 7 presents two \textit{IuST}s searched for the path containing “Current sea level rise” of the relationship between “Global warming” and “Agriculture.” The \textit{IuST} of ID 1 is searched using the Top-Down approach, which includes a little knowledge about the fact that global warming affects agriculture through rising sea levels. The \textit{IuST} of ID 1 is searched because its surrounding text contains several elucidatory objects of the relationship between “Global warming” and “Agriculture,” including “Current sea level rise.” The \textit{IuST} of ID 2 is searched using the Bottom-Up approach, which includes detailed knowledge of the relationship between “Global warming” and “Current sea level rise.” As presented in column “Elucida-
tory objects” in Fig. 7, the surrounding text of the \textit{IuST} of ID 2 contains many elucidatory objects of the relationship between “Global warming” and “Current sea level rise.” Testers consider the \textit{IuST} of ID 2 include more knowledge about the path, and assign a score 2 to it. Therefore, we conclude that the Bottom-Up approach is more appropriate to search \textit{IuST}s for a path of a relationship.

To assess the reliability of the agreement between the 10 testers, we compute the Fleiss’ kappa between the scores given by each tester for the \textit{IuST}s obtained using each approach. Table 2 presents the \( k \) values of the Fleiss’ kappa for each approach. If the testers are in complete agreement, then \( k = 1 \); if there is no agreement among the testers then \( k = 0 \). According to the interpretation of the \( k \) values given by Landis and Koch [13], \( 0.01 \leq k \leq 0.20 \) indicate a “slight agreement,” \( 0.21 \leq k \leq 0.40 \) indicate a “fair agreement” and a “Moderate agreement,” respectively. Therefore, we conclude that the 10 testers achieved a fair agreement on the evaluation for each approach, and the results depicted in Fig. 5 and Fig. 6 are reliable.

We implement Enishi using the Top-Down and the Bottom-Up approach to search \textit{IuST}s for the overall relationship and each path of a relationship, respectively.

4.2.2 Discussion about Query Extension

Several query extension methods [14], [15] were proposed to search results of high precision about an object. It was also natural to consider to search \textit{IuST}s about the relationship between \( s \) and \( t \) using a keyword image search engine with the query “\( st_0 \)” that is extended from “\( st \)” with one of the elucidatory objects \( o_i \). However, there are so many elucidatory objects for a relationship. For example, there are 123 elucidatory objects extracted from Wikipedia for the relationship between “Global warming” and “Agriculture.” It is unobvious that which elucidatory objects are effective for constructing queries to search \textit{IuST}s that include a lot of knowledge about a relationship. Therefore, we do not adopt the query extension approach.

We also conduct a case study to discuss query extension compared with our approach in the following. Let
In most cases, other pages on average. It is possible to mine at least 100 using the top-100 paths. Therefore, we set ing the top-150 paths, are almost identical to those searched (2010 evidence-based method. In the English Wikipedia dataset although the increase becomes slow when time when the number

We examine the eff ect of elucidatory objects on the evidence-based method. For a relationship, we obtain the top-300 IuSTs using Google; we then count the number \( a(k) \) of IuSTs whose text segment contains at least one elucidatory object in the top-\( k \) paths, where \( 1 \leq k \leq 150 \). Figure 9 plots the average number \( a(k) \) of the 40 relationships for each \( k \). The number \( a(k) \) is monotonically increasing, although the increase becomes slow when \( k \) becomes larger. For example, the difference 28 between \( a(100) = 186 \) and \( a(50) = 158 \) is fairly large, while \( a(150) - a(100) \) is only 12. On the other hand, the evidence-based method costs much time when the number \( k \) of used paths is large. In preliminary experiments, we also observed that IuSTs searched using the top-150 paths, are almost identical to those searched using the top-100 paths. Therefore, we set \( k = 100 \) for the evidence-based method. In the English Wikipedia dataset (2010/03/12 snapshot), a page has 40.25 links linking to other pages on average. It is possible to mine at least 100 paths important for a relationship on the Wikipedia dataset in most cases.

4.2.3 Observation of Elucidatory Objects

We also compute the percentage \( p(k) \) of elucidatory objects in the top-\( k \) paths appearing in any of the top-300 IuSTs retrieved using Google, for a relationship. Figure 10 plots the average \( p(k) \) of the 40 relationships for each \( k \). 58%, 44%, and 41% of the elucidatory objects in the top-10, top-50, and top-100 paths on average for the 40 relationships appear in at least one IuST, respectively. The average \( p(k) \) decreases along with the increase of \( k \). On average, for the 40 relationships, the number of elucidatory objects in the top-10, top-50, top-100, and top-150 paths are 13, 67, 145, and 220. Therefore, a fairly large number of elucidatory objects are actually used in the evidence-based method. Furthermore, elucidatory objects in highly ranked paths are more effective than those in lowly ranked paths.

4.2.4 Novelty of Knowledge Retrieved from the Web

In Enishi, we show to users the top-10 paths mined from Wikipedia for a relationship. To complement the knowledge about a relationship represented by the 10 paths, we also retrieve IuSTs from the Web. We assume users could obtain knowledge from the retrieved IuSTs which is unavailable by reading the top-10 paths and by reading Wikipedia pages corresponding to the objects in the top-10 paths. In this section, we evaluate the novelty of the knowledge retrieved from the Web compared to that existing in Wikipedia.

For a relationship, we extract a set \( N_{Wiki} \) of nouns from Wikipedia pages of the objects in the top-10 paths mined for
the relationship; and a set $N_{Web}$ of nouns from the surrounding text of the $IwST$s searched for the overall relationship using the Top-Down approach. We use a part-of-speech tagger\footnote{http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/} to extract nouns from text. We consider that the knowledge about the relationship searched on the Web is not contained in Wikipedia if many nouns in $N_{Web}$ are not contained in $N_{Wiki}$. For every relationship, we obtain two sets of $N_{Web}$, denoted by $N_{Web}(T10)$ and $N_{Web}(>0)$, respectively from the top-10 $IwST$s and all the $IwST$s having positive scores. Figure 11 depicts the rates of nouns appearing in $N_{Web}$ but not in $N_{Wiki}$ for each relationship. On average, for all 40 relationships, 50.4% nouns in $N_{Web}(T10)$, and 61.2% nouns in $N_{Web}(>0)$, do not belong to $N_{Wiki}$. The average number of nouns in $N_{Wiki}$, $N_{Web}(T10)$ and $N_{Web}(>0)$ are 5990, 3080 and 4798, respectively. Therefore, there is a possibility that the $IwST$s retrieved from the Web include knowledge which is unavailable from the Wikipedia pages corresponding to the objects in the top-10 paths.

We also actually confirmed that the surrounding text includes knowledge unavailable in Wikipedia. Figure 12 presents the top-5 $IwST$s for the overall relationship between “Global warming” and “Agriculture,” searched using the Top-Down approach. The surrounding text of the second image consists of 77 sentences, some of which are presented in the “Surrounding text extracted” column. The surrounding text contains 299 nouns, and about 24.4% of the 299 nouns do not appear in the $N_{Wiki}$ for the relationship. We checked every sentence carefully for whether the knowledge represented by the sentence is contained in Wikipedia. We found that 69 of the 77 sentences include knowledge unavailable from Wikipedia, although only 24.4% of the nouns in the text do not appear in Wikipedia. Especially, the text contains many precise data that are not described in Wikipedia. Such as, “Global warming is exposing an additional 69 million to 91 million people to food shortages by 2085;” “Due to development and agriculture, the forested area of the world is expected to fall 25 percent to 30 percent by 2050 and the area of coastal wetlands are expected to decline 40 percent by 2085;” and “Between now and 2085, global warming could increase forested areas by 5 percent; but it could reduce the area of coastal wetlands another 13 percent.” Similarly, we also find many other surrounding text segments retrieved from the Web include knowledge unavailable from Wikipedia.
4.3 Case Study: Relationship between Global Warming and Agriculture

Figure 12 presents the top-5 *IuSTs* for the overall relationship between “Global warming” and “Agriculture,” searched using the Top-Down approach. The “Surrounding text extracted” column presents some sentences in the surrounding text of each image extracted using our program. The “Elucidatory objects” column presents some elucidatory objects appearing in the surrounding text of each *IuST*. The numbers in the “score” column are the average scores obtained by a human subject. As discussed in Sect. 2, an image itself is insufficient for understanding a relationship. However, reading the images with their surrounding text, users could understand the relationship. By reading the summaries of surrounding text, we ascertained that each image is relevant to its surrounding text and all the five *IuSTs* include knowledge about the relationship between “Global warming” and “Agriculture.” As listed in the “Elucidatory objects” column, many elucidatory objects playing important roles in the relationship, appear in the surrounding text of these *IuSTs*. Especially, *IuST* ranked 2nd, 3rd, and 4th include a lot of knowledge about the relationship. The image of the *IuST* ranked 1st does not relate to the relationship directly, although its surrounding text includes knowledge about the relationship. The *IuST* ranked 5th presents a farming technique that is intended to solve the greenhouse gas problem. The two *IuSTs* ranked 1st and 5th include less knowledge about how global warming and agriculture affect each other than the other three *IuSTs* do. Therefore, testers assigned lower scores for the two *IuSTs*. The “Google” column presents rankings generated by the Google image search engine using a query “Global warming Agriculture.” The images ranked 2nd and 4th using our approach are ranked very low by Google, although the two include a wealth of knowledge about the relationship. It is difficult for users to search these images themselves using a keyword search engine. Our approach can search these images for the users.

5. Related Work

In this paper, we described a new method to search *IuSTs* that include knowledge about a relationship on the Web by utilizing the elucidatory objects mined from Wikipedia. Our work would be an information retrieval research using data mining techniques. Several query extension methods also use data mining techniques for searching information. Taneva et al. [14] search photos of objects, such as people, building, or mountains. They first mine objects related to a certain object from a semantic knowledge base. They then search photos of an object using a keyword image search engine with queries extended with the mined related objects. For example, searching photos of a mountain uses queries constituted by its name and its location or height. Similar to the methods, we search knowledge on the Web using knowledge mined from Wikipedia. However, we search knowledge about relationships between objects rather than that about single objects. As another related work, KORU [15] is a search engine, which extends queries using Wikipedia to help users to search what they did not know how to search. Our method is related to but different from these query extension methods. As confirmed in Sect. 4.2.2, the top-10 *IuSTs* searched for a relationship between objects and using our method can not be obtained sufficiently by extending queries of “s t” with one elucidatory object.

To the best of our knowledge, no method was proposed for searching Web pages or images including knowledge about complicated implicit relationships on the Web. However, many methods exist for extracting explicit relationships from Web pages, such as Wikipedia articles. The DBpedia project [5] extracts semantic relationships between objects from the infobox in Wikipedia articles. For example, the relationship that “Tokyo” is the “capital” of “Japan,” could be obtained from the infobox of article “Japan.” Lehmann et al. [16] then developed a tool for exploring relationships extracted by DBpedia. YAGO [4], [17] is a semantic knowledge base extracted from Wikipedia and WordNet. YAGO contains about two million objects and about 20 million binary relationships about them, such as “Politician” is a subclass of “person”, and “Max Planck” has won the “Nobel Prize”. Kasneci et al. [18] then proposed a search engine to search semantic relationships between two objects using YAGO. Both DBpedia and YAGO do not extract relationships represented by the links existing in the body text of Wikipedia articles. Many relationships existing in the Wikipedia information network are not included in DBpedia or YAGO.

Zhu et al. [9] extract semantic relationships between pairs of people from the Web using natural language processing technologies. Bollegala et al. [7] proposed a method to extract semantic relationships between any objects from the Web, such as X was born in Y, X is a CEO of Y, and X buys Y. As discussed in Sect. 1, seeking semantic relationships is insufficient for understanding complicated relationships in the real world. EntityCube [9] and SPYSEE [8] extract explicit relationships between pairs of people from the Web. They then search people having explicit relationship to a given people. Both EntityCube and SPYSEE do not search knowledge explaining an implicit relationship between two people.

6. Conclusion

We proposed approaches to search images with surrounding text that include knowledge about a relationship on the Web, by utilizing objects constituting the relationship mined from Wikipedia. We performed experiments to evaluate our approaches with human subjects and by comparing them with Google image search engine. Experimental results revealed that our approaches are able to search images with surrounding text that include knowledge about relationships. Furthermore, the information retrieved from the Web include new knowledge unavailable in Wikipedia.
We also construct a relationship search system named Enishi, based on the proposed approaches. Enishi is an applicable system for users to search diverse knowledge from Wikipedia and the Web to understand relationships deeply.

References

[1] X. Zhang, Y. Asano, and M. Yoshikawa, “Enishi: Searching knowledge about relations by complementarily utilizing Wikipedia and the web,” Proc. 11st WISE, pp.480–495, 2010.

[2] X. Zhang, Y. Asano, and M. Yoshikawa, “A generalized flow based method for analysis of implicit relationships on Wikipedia,” IEEE Trans. Knowl. Data Eng., vol.25, no.2, pp.246–259, 2013.

[3] X. Zhang, Y. Asano, and M. Yoshikawa, “Mining and explaining relationships in Wikipedia,” IEICE Trans. Inf. & Syst., vol.E95-D, no.7, pp.1918–1931, July 2012.

[4] F.M. Suchanek, G. Kasneci, and G. Weikum, “Yago: A core of semantic knowledge,” Proc. 16th WWW, pp.697–706, 2007.

[5] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, “DBpedia: A nucleus for a web of open data,” in The Semantic Web, ed. K. Aberer, K.S. Choi, N. Noy, D. Allemang, K.I. Lee, L. Nixon, J. Golbeck, P. Mika, D. Maynard, R. Mizoguchi, G. Schreiber, and P. Cudre-Mauroux, pp.722–735, Springer, 2007.

[6] G.R. Barnes, P.B. Cerrito, and I. Levi, “Construction of well separated interaction semigroups to refine the modeling of social networks,” Social Networks, vol.25, no.2, pp.161–174, 2003.

[7] D. Bollegala, Y. Matsuo, and M. Ishizuka, “Relational duality: unsupervised extraction of semantic relations between entities on the web,” Proc. 19th WWW, pp.151–160, 2010.

[8] SPYSEE. http://spysee.jp.

[9] J. Zhu, Z. Nie, X. Liu, B. Zhang, and J.R. Wen, “Statsnowball: a statistical approach to extracting entity relationships,” Proc. 18th WWW, pp.101–110, 2009.

[10] Z. Chen, L. Wenyin, F. Zhang, M. Li, and H. Zhang, “Web mining for web image retrieval,” JASIST, vol.52, no.10, pp.831–839, 2001.

[11] R. Datta, D. Joshi, J. Li, and J.Z. Wang, “Image retrieval: Ideas, influences, and trends of the new age,” ACM Computing Surveys, vol.40, no.2, 2008.

[12] F. Fauzi, J.L. Hong, and M. Belkhair, “Webpage segmentation for extracting images and their surrounding contextual information,” Proc. 17th ACM International Conference on Multimedia, pp.649–652, 2009.

[13] J.R. Landis and G.G. Koch, “The measurement of observer agreement for categorical data,” Biometrics, vol.33, no.1, pp.159–174, 1977.

[14] B. Taneva, M. Kacimi, and G. Weikum, “Gathering and ranking photos of named entities with high precision, high recall, and diversity,” Proc. 3rd WSDM, pp.431–440, ACM, 2010.

[15] D.N. Milne, I.H. Witten, and D.M. Nichols, “A knowledge-based search engine powered by Wikipedia,” Proc. 16th CIKM, New York, NY, USA, pp.445–454, 2007.

[16] J. Lehmann, J. Schüppel, and S. Auer, “Discovering unknown connections - the DBpedia relationship finder,” Proc. 1st Conference on Social Semantic Web, pp.99–110, 2007.

[17] G. Weikum, G. Kasneci, M. Ramanath, and F. Suchanek, “Database and information-retrieval methods for knowledge discovery,” Commun. ACM, vol.52, no.4, pp.56–64, 2009.

[18] G. Kasneci, F.M. Suchanek, G. Ifrim, M. Ramanath, and G. Weikum, “Naga: Searching and ranking knowledge,” Proc. 24th ICDE, pp.953–962, 2008.

[19] K.D. Wayne, Generalized Maximum Flow Algorithm, Ph.D. thesis, Cornell University, New York, U.S., 1999.

[20] K. Nakayama, T. Hara, and S. Nishio, “Wikipedia mining for an association web thesaurus construction,” Proc. 8th WISE, pp.322–334, 2007.

[21] Y. Koren, S.C. North, and C. Volinsky, “Measuring and extracting proximity in networks.” Proc. 12th ACM SIGKDD Conference, pp.245–255, 2006.

Appendix: Methods for Mining Knowledge of Relationships in Wikipedia

In the Appendix, we introduce the method for mining disjoint paths explaining a relationship in Wikipedia, as proposed by Zhang et al. [3]. Before that, we first explain its basis: the generalized max-flow model proposed by Zhang et al. [2] for measuring relationships.

A.1 Generalized Max-Flow Model for Measuring Relationships

Zhang et al. [2] model a relationship between two objects in a Wikipedia information network using a generalized max-flow. The generalized max-flow problem [19] is identical to the classical max-flow problem except that every edge $e$ has a gain $\gamma(e) > 0$; the value of a flow sent along edge $e$ is multiplied by $\gamma(e)$. Let $f(e) \geq 0$ be the amount of flow $f$ on edge $e$, and $\mu(e) \geq 0$ be the capacity of edge $e$. The capacity constraint $f(e) \leq \mu(e)$ must hold for every edge $e$. The goal of the problem is to send a flow emanating from the source into the destination to the greatest extent possible, subject to the capacity constraints. Let generalized network $G = (V, E, s, t, \mu, \gamma)$ be information network $(V, E)$ with the source $s \in V$, the destination $t \in V$, the capacity function $\mu$, and the gain function $\gamma$. Figure A-1 depicts an example of a generalized max-flow. 0.4 units and 0.2 units of the flow arrive at “USA” along path (A) and path (B), respectively. As illustrated in Fig. A-1, a large amount of flow is usually sent along disjoint paths which are short and are formed by edges having high gains.

To measure the strength of a relationship from object $s$ to object $t$, Zhang et al. [2] use the value of a generalized maximum flow emanating from $s$ as the source into $t$ as the destination; a larger value signifies a stronger relationship. Such a flow seldom uses an edge whose direction is opposite that from the source to the destination; a larger value signifies a stronger relationship.

![Fig. A-1](image-url) A generalized max-flow and its decomposition.

\[ \text{Fig. A-1} \] A generalized max-flow and its decomposition.
important for a relationship in Wikipedia, such as the path \((\text{Global warming, Carbon dioxide, Agriculture})\) discussed in Sect. 2.3. To use paths formed by edges of any direction, Zhang et al. construct a doubled network \([2]\) by adding to every original edge a reversed edge whose direction is opposite to the original one.

A path formed by edges representing important explicit relationships in constituting an implicit relationship, are usually important to the implicit relationship. An implicit relationship represented by many of such paths is usually a strong relationship. Therefore, it is desired to assign larger gains to edges that are important in constituting a relationship in the model. To realize such a gain assignment, Zhang et al. \([2]\) proposed an edge gain function using the category structure of Wikipedia. We omit details of the gain function here because of space limitations.

The model reflects all the three concepts important for measuring a relationship: distance, connectivity and co-citation. Zhang et al. \([2]\) ascertained the model can measure the strength of relationships more correctly than previous methods \([20],[21]\) can.

A.2 Generalized Flow Based Method for Mining Disjoint Paths

Based on the generalized max-flow model, Zhang et al. proposed a method to mine disjoint paths important for understanding a relationship from object \(s\) to object \(t\) in Wikipedia \([3]\), by extracting paths along which a large amount of a generalized max flow is sent from \(s\) to \(t\).

We introduce the method presented below. (1) Construct a generalized network \(G = (V,E,s,t,\mu,\gamma)\) using pages and links within, at most, \(m\) hop links from \(s\) or \(t\) in Wikipedia. (2) Construct the doubled network \(G'\) for \(G\), determine the edge gain \(\gamma\). (3) Compute a generalized max-flow \(f\) emanating from \(s\) into \(t\) on \(G'\). (4) Decompose the flow \(f\) into flows on a set \(P\) of paths. Let \(\gamma(p_i)\) denote the value of flow on a path \(p_i\), for \(i = 1, 2, \ldots, |P|\). For example, the flow on the network depicted in Fig. A.1 is decomposed into flows on two paths (A) and (B). The value of the decomposed flow on path (A) is 0.4; that on path (B) is 0.2. (5) Output the top-\(k\) paths in decreasing order of \(\gamma(p_i)\).

As explained in Sect. A.1, in the generalized max-flow model, a large amount of flow is usually sent along paths which are short and are formed by edges having high gains, such as, in Fig. A.1, the flow is sent along path (A) and path (B), and the value of the flow on path (A) is larger than that on path (B). Consequently, the top-\(k\) paths that hold high values of flows, usually are short and are formed by edges having high gains. Zhang et al. \([3]\) assign a larger gain to edges that are important in constituting a relationship. Therefore, the top-\(k\) paths usually are short and are formed by important edges. Generally, short paths are more important for a relationship than long paths, as discussed in \([21]\). Also, there is a high possibility that paths formed by important edges are important. Therefore, a high possibility exists that the top-\(k\) paths are important to a relationship.

Zhang et al. \([3]\) confirmed through experiments that the generalized flow based method can mine many disjoint paths useful for understanding a relationship. For example, Fig. 3 depicts the largest-10 paths mined for the relationship between “Global warming” and “Agriculture,” using the method.