Discovery of Surrounding Fact Information Based on Fact Adjacency Relationships

Meng ZHAO[†], Hiroaki OHSHIMA[†], and Katsumi TANAKA[†]

[†] Department of Social Informatics, Graduate School of Informatics, Kyoto University
Yoshida Honmachi, Kyoto, 606–8501 Japan
E-mail: [†]{zhao,ohshima,tanaka}@dl.kuis.kyoto-u.ac.jp

Abstract During daily web surfing, users encounter vast quantities of information everyday and at most time just pass by. It maybe because there is no more time for further search, or even users did not notice the information at all. However, among those missed information, there are some meaningful pieces. In this paper, we propose a system that given a fact, surrounding information of the fact are discovered, and then ranked according to cognition difference with context formed by surrounding facts. The objective of our system is to help users efficiently find related and useful information, reducing users ’ searching cost.

Key words adjacency relationships, context cognition

1. Introduction

Nowadays, web has become the primary access to acquire various information. Everyday we encounter vast quantities of information, but only a small portion of them is viewed by us. This does not mean the rest is insignificant. The reasons we miss those significant information may be because even though we realize their importance, there is no more time for deep search. Or we are not aware of them at all. Suppose during daily web surfing, the following information is read by chance.

Lemons are rich in Vitamin C.

Depending on one's next move, users can be divided into two types: those who stop there and those who regard it as a starting point and begin to try finding more information. Information such as the following ones are considered to be meaningful for the latter.

1). *Lemons are rich in Citric acid.*

2). *Acerola is known for being extremely rich in Vitamin C.*

3). *Citrus fruits are full of Vitamin C.*

4). *Fresh food is the best source of Vitamin C and most fruits and vegetables contain Vitamin C.*

5). *Lemons contain no more Vitamin C compared to acerolas.*

6). *Citrus fruits are full of fiber and important minerals.*

Take an insight into the above enumerative information. If one is interested in *lemon*, then 1). and 5). meet his need. Or if one is interested in *Vitamin C*, then 2)., 3)., 4). and 5). meet his need. Especially, 3). and 4). state in a more general way, while 5). shows the comparison of the Vitamin C amount between lemon and acerola. In a broader sense, 6). also provides useful information about *lemon*, since we know lemon is a kind of citrus fruits so that we can easily infer that *"Lemons are full of fiber and important minerals"*. Besides, it does not mean that such kind of information is meaningless for those who stop further search. Perhaps they just cannot catch its significance as an entrance to the wider knowledge environment which leads to their unawareness. Thus, for both those who stop there and those who regard *"Lemons are rich in Vitamin C"* as a starting point and begin to try finding more information, information like the above listed in the example helps them acquire more useful and related knowledge indeed.

From the above discussion, we summarize three distinct relationships: **variable**, **generalization** · **specialization**, **conditional negation** · **counterevidence**. (See details in Section 4.) All of them are referred to as adjacency relationships here. Briefly speaking, **variable** corresponds to the situation that an entity is substituted by another entity. In the above example, "Citric acid" substitutes "Vitamin C" in 1). **Generalization** · **specialization** corresponds to the situation that an entity is substituted by its hypernym or hyponym, i.e. "citrus fruits" substitutes "lemons" in 3). **Conditional negation** · **counterevidence** corresponds to the situation that contrast and contradiction occur, i.e. 5).

In this paper, we take a sentence fact as a unit and propose a method to find its surrounding facts in accordance with the distinct relationships mentioned before. [注1] Besides, we also introduce a context-aware ranking method to rank them so as to present more important facts in the upper.

The remainder of the paper is organized as follows. In Section 2., we discuss related work. Section 3. shows the overall workflow of this work, while Section 4. states the details of adjacency relationships between facts. In Section 5., we introduce the method to generate candidate facts by using dictionary and extract valid ones from the Web. In Section 6., our context-aware ranking method is illustrated. Finally, Section 7. concludes the paper and gives an outline of our future work.

2. Related Work

2. 1 Contradiction detection in nature language processing

In nature language processing field, some work focuses on detecting contradiction between two sentences. In [3], Harabagiu et al. concentrate on three forms of linguistic information - negation, antonymy, semantic and pragmatic information, employ a machine learning approach and provide the first empirical results for contradiction detection. However, Marneffe et al. [1] advocate that contradictions are not limited to these constructions and broader extent should be covered. As a result, they propose 9 types for contradiction and categorize them as those occurring via antonymy, negation, date/number mismatch and those arising from the use of factive words, structural and subtle lexical contrasts, and world knowledge. Although some types of contradiction are beyond their system's capability, they get good performance on types arising from negation and antonymy. Given a topic term, WISDOM system [注2] developed by National Institute of Information and Communications Technology (NICT) [注3], can extract contrast sentences from a set of documents without any training data. This is also the essential difference compared to [1] and [3]: the objective of WISDOM system is to find related keywords of a given topic and present primary, contrast sentences to users, while [1] and [3] aim at detecting relations between two sentences. All of them are based on analysis of part-of-speech (POS) tagging.

2. 2 Credibility judgment

Although one can get a vast amount of information from the Web, there mix some unreliable information. To assist users in easily distinguishing reliable information from unreliable one, some work has been done to judge the credibility of a given sentence. In [7] and [8], Yamamoto et al. propose a system called "Hondo? Search" to help users determine the trustworthiness of uncertain facts based on sentiment and temporal viewpoints. With the help of some additional data, users can judge an uncertain fact by themselves. Unreliability arising from temporal changes is also taken account of. In [5], they improve Hondo? Search to not only collect comparative facts of the input fact, but only provide users important aspects for comparison. Especially, in addition to an uncertain fact whose credibility one wants to check, a verification target should also be indicated by user. For example, "Lemons are rich in Vitamin C" as the uncertain fact and "Lemons" as the verification target. In these three work, lexical-syntactic patterns of the sentence substitutions are used to extract comparative facts from the Web. Compared with our work, we apply lexical-syntactic patterns to get both comparative facts and conditionally negational ones. Besides, dictionaries containing hypernym/hyponym and coordinate terms are also used to generate candidate related facts.

2. 3 Search by sentence queries

In [6], Yamamoto and Tanaka concentrate on improving search results responded by sentence queries. They advocate two cases in which satisfied search results cannot be obtained by sentence queries: (1) the meaning of sentence is correct but its expression is rare on the Web; (2) what the sentence describes is a misunderstanding by user. In their example, given a sentence query, i.e. "Germany is famous for beer", phrases such as "Belgium is famous for beer", "Munich is famous for beer" are suggested as substitutions of input query. Based on the criteria that sentence substitutions which appears frequently on the Web and whose context is similar to that of the input sentence query should be ranked higher, a ranking algorithm is also stated. Our **variable** is similar to [6] in this respect. However, we also consider sentence substitution based on hypernym/hyponym. Besides, our ranking algorithm is context-aware, which means a sentence's rank depends on users' acceptability of other sentences.

3. Overview

In this section, overview of our work is described. We aim at discovering surrounding fact information of a user-given fact, while the surrounding is determined by three adjacency relationships - variable, generalization · specialization, conditional negation · counterevidence. We believe with the help of those information, users are able to easily and efficiently obtain meaningful pieces of information of their input fact. Figure 1 shows the overview by using an example fact *"Lemons are rich in Vitamin C"*.

The one and only input for our proposed algorithm is a fact

[注1] : In this paper, we only deal with the former two relationships.

[注2] : http://kc.nict.go.jp/project1/WISDOM_TR.pdf

[注3] : http://www.nict.go.jp/en/index.html

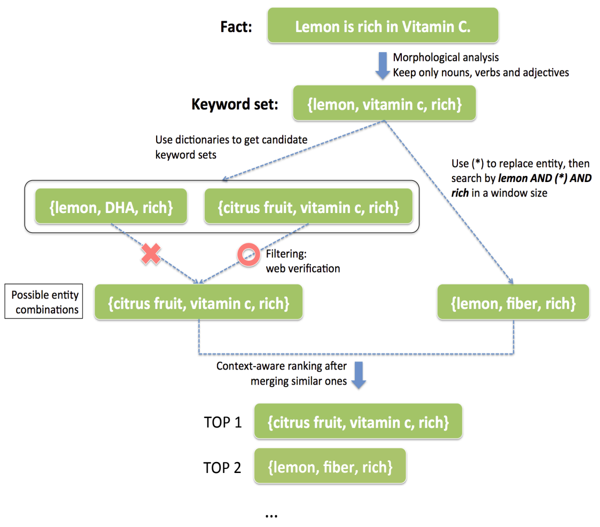
Fact:
Lemon is rich in Vitamin C.
Morphological analysis
Keep only nouns, verbs and adjectives
Keyword set:
{lemon, vitamin c, rich}
Use dictionaries to get candidate keyword sets
Use (*) to replace entity, then search by lemon AND (*) AND rich in a window size
{lemon, DHA, rich}
{citrus fruit, vitamin c, rich}
Filtering: web verification
Possible entity combinations
{citrus fruit, vitamin c, rich}
{lemon, fiber, rich}
Context-aware ranking after merging similar ones
TOP 1
{citrus fruit, vitamin c, rich}
TOP 2
{lemon, fiber, rich}
...

Figure 1 Overview.

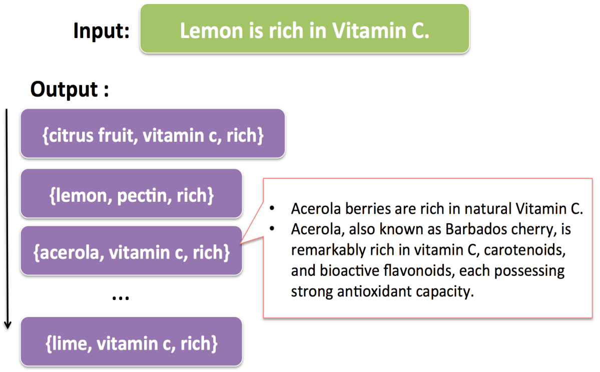
Input:
Lemon is rich in Vitamin C.
Output :
{citrus fruit, vitamin c, rich}
{lemon, pectin, rich}
{acerola, vitamin c, rich}
...
{lime, vitamin c, rich}
• Acerola berries are rich in natural Vitamin C.
• Acerola, also known as Barbados cherry, is remarkably rich in vitamin C, carotenoids, and bioactive flavonoids, each possessing strong antioxidant capacity.

Figure 2 An example of input and output.

one wants to know more about. In general, if one wants to know more about a fact, it is necessary for him to search by this fact and check search engine's response. And not least, he may need to change his queries several times in order to get what he desires for. Here a fact is input as a sentence, such as *"Lemons are rich in Vitamin C"* shown in Figure 1. After a fact is input, at the very beginning, morphological analysis is taken to get a keyword set from it, i.e. {lemon, vitamin c, rich}. Then according to the source from where a surrounding fact candidate is obtained, two different ways are employed:

a). Using dictionary information such as hypernym/hyponym to generate surrounding fact candidates from the input fact.

b). Applying web search to get surrounding fact candidates.

For the former case, incorrect combinations may arise because of simple hypernym/hyponym substitution. For example, since "DHA" is a coordinate term of "Vitamin C", a substitution can be generated as {lemon, DHA, rich} which means lemons are rich in DHA. But in fact, lemons contain no DHA at all. That is to say, entity combination of "lemon" and "DHA" is impossible. Thus, it should be removed before fact keyword set ranking. Besides, similar keyword sets are merged. Finally, in order to show users more meaningful and important surrounding facts in the upper, keyword sets obtained from the above steps are ranked by considering their context [注4]. We also provide the details of each keyword set for users' further reading. Figure 2 shows an output image of the input fact *"Lemon is rich in Vitamin C"*. We can see there return some keyword sets, such as the top one is {citrus fruit, vitamin c, rich}, the second one is {lemon, pectin,rich}. Moreover, concentrated on the keyword set {acelora, vitamin c, rich}, we can also find that its detailed facts such as *"Acerola berries are rich in natural Vitamin C"* and *"Acerola, also known as Barbados cherry, is remarkably rich in vitamin C, carotenoids, and bioactive flavonoids, each possessing strong antioxidant capacity"* are also attached.

4. Fact Adjacency Relationships

According to the Oxford Dictionary [注5], a *fact* is a thing that is known or proved to be true. Based on this definition, we hypothesize a fact should meet either the following conditions: (1) What the fact states is accepted by most people. In other words, it appears frequently on the Web. (2) What the fact means is true. We apply three distinct relationships between facts to obtain surrounding facts of a given one. Details for each relationship are discussed as follows.

4.1 Variable

Given a fact, we suppose that its surrounding facts can be generated by replacing an entity. Here an entity indicates a noun keyword extracted from the given fact. For example, in the fact f = *"Lemons are rich in Vitamin C"*, there are two entities - "lemon" and "vitamin c". Thus, fact *"Acelora berries are rich in Vitamin C"* has the same lexico-syntactic pattern as that of f (*X are rich in Vitamin C*), while fact *"Lemons are rich in fiber"* also has the same lexico-syntactic pattern as that of f (*Lemons are rich in X*). We regard this kind of substitutions of an entity as "variable".

4.2 Generalization · Specialization

Given a fact, we suppose that facts generalized or specialized the meaning it has are also its surrounding facts. Take the fact f = *"Citrus fruits are rich in Vitamin C"* as an example. For the term "vitamin c", the term "vitamin" is one of its hypernym, as Vitamin C is a kind of Vitamin. Therefore, fact *"Citrus fruits are rich in Vitamin"* is a generalization of

[注4]：What we say "context" here does not indicate the paragraph or document from which the fact keyword set is extracted.

[注5]：http://www.oxforddictionaries.com

f, since it provides a broader description about citrus fruits to users. On the other hand, for the term "citrus fruit", the term "lemon" is one of its hyponym, as lemon is one of the citrus fruits. Therefore, fact *"Lemons are rich in Vitamin C"* is a specialization of f, since it gives users a detailed example of citrus fruits. We regard these kinds of substitutions of an entity as "generalization · specialization".

4. 3 Conditional negation · Counterevidence

Given a fact, we suppose its surrounding facts should be also included those who has a negative or contrastive implication of it. For example, when the fact $f =$ *"Lemons are rich in Vitamin C"* is shown to users, they can catch the meaning - rich Vitamin C contained in lemons. However, it is hard to know its rich degree. If fact *"Lemons contain no more Vitamin C than acelora berries"* is presented, then users are able to have a more clear understanding about lemons in the viewpoint of Vitamin C content. Since simple negation of a fact, i.e. *"Lemons are not rich in Vitamin C"*, is always incorrect, we only consider negation of a fact with some conditions, say "conditional negation". Besides, it is likely that the input fact is actually a misunderstanding by users. In this case, presentation of counterevidences is considered to be essential. We regard the above surrounding as "conditional negation · counterevidence".

5. Finding Surrounding Fact Keyword Sets

Given a fact f, we wish to present its surrounding facts s of f by adjacency relationships. We define this as $f\Theta s$. The goal in this section is to gather surrounding fact keyword sets $S = \{s|f\Theta s\}$. As we mentioned in Section 3., depending on the source from where a surrounding fact is obtained, two different ways are employed. See below.

5. 1 Dictionary-based approach

Given a fact f, we first refine it to a keyword set K_f. The keywords are the ones which consist of the fact and are not stopwords. Then according to lexical category of each keyword, substitutions using dictionary information are taken place. At this time, incorrect combinations may arise because of simple substitutions so that a verification process is conducted to filter all possible combinations. For example, given a fact $f =$ *"Lemons are rich in Vitamin C"*, its corresponding keyword set is {lemon, vitamin c, rich}. Since term "citrus fruit" is a hypernym of "lemon", a substitution is {citrus fruit, vitamin c, rich}. Similarly, since term "DHA" is a coordinate term of "vitamin c", a substitution is {lemon, DHA, rich}. However, *"lemons are rich in DHA"* is not true, so we remove the substitution {lemon, DHA, rich}. The substitution {citrus fruit, vitamin c, rich} together with other possible combinations are kept for ranking.

We use WordNet [注6] to generate candidate keyword sets of fact f and verify their possibilities as follows:

（1） The given fact f is divided into words by taking morphological analysis. Then according to defined stopword list, some words are omitted. The remaining words are denoted as $K_f = \{k_1, k_2, ..., k_n\}$ and we call K_f the *keyword set* of fact f.

（2） For keyword k_i in K_f, if it is a noun, replace it by its direct hypernym, direct hyponym, or sister term in WordNet. If it is a verb, replace it by its sister term. If it is an adjective, since in this paper we skip conditional negation · counterevidence relationship, only replace it by adjectives which is semantically similar to it, in other words, "sister term" in WordNet.

（3） Verify the possibility of keyword combination by querying the Web. Given a substitution keyword set $K_s = \{k'_1, k_2, ..., k'_n\}$, where k'_i indicates a replacement of k_i from K_f by using WordNet. Its combination possibility is calculated as $p(K_s) = \frac{HitCount(k'_1 \wedge k_2 \wedge ... \wedge k'_n)}{\frac{1}{{}_rC_{r-1}} HitCount(Q_s)}$, where $HitCount(k'_1 \wedge k_2 \wedge ... \wedge k'_n)$ is the hitcount by issuing query "$k'_1 \wedge k_2 \wedge ... \wedge k'_n$" to the Web, Q_s indicates the query set $Q_s = \{k_2 \wedge ... \wedge k_{n-1}, ..., k'_1 \wedge ... \wedge k'_n\}$ and r is the number of replaced keywords compared to K_f. Actually, here we aim at estimating the possibility of the combination of the r replacements. For example, if $K_f =$ {lemon, vitamin c, rich} and $K_s =$ {lemon, DHA, rich}, then for this K_s, its $Q_s =$ {lemon $\wedge$ rich, DHA $\wedge$ rich}. We check the possibility of the combination of "lemon" and "DHA". Only K_s whose combination possibility is above a threshold is remained.

5. 2 Web-search-based approach

Given a fact f, we first refine it to a keyword set K_f. Then according to morphological analysis, only nouns (regarded as entities) are selected for variable transformation. We search the Web to gather candidate surrounding keyword sets K_s as follows:

（1） Morphological analysis is applied to the input fact f. Only nouns, verbs and adjectives are extracted. The same as what we state before, we denote the extracted keywords as $K_f = \{k_1, k_2, ..., k_n\}$.

（2） For each noun in K_f, we regard it as an entity. Documents denoted as $Doc(K_s)$ which may contain surrounding facts of f are extracted by issuing query "$(*) \wedge k_2 \wedge ... \wedge k_n$" to a conventional Web search engine. In this example, k_1 is an entity but maybe not the only one. To simplify the problem, we only consider one entity replacement by asterisk at

[注6]：http://wordnet.princeton.edu
WordNet is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms. Besides, hypernym/hyponym of nouns are also available.

a time in a query. That means even there are two entities in K_f ={lemon, vitamin c, rich}, we only consider the following two queries: "(*) ∧ vitamin c ∧ rich" and "lemon ∧ (*) ∧ rich".

（3） Sentences in $Doc(K_s)$ which contain all keywords [注7] in the above queries in a window size are extracted. Then for each sentence, a surrounding keyword set K_s is refined.

Similar K_s obtained by dictionary-based approach and Web-search-based approach are merged. For example, a K_s ={rich, vitamin c, lemon, health, benefit} refined from the fact f = *"Rich in Vitamin C, lemon has many health benefits"*, is merged as {lemon, vitamin c, rich}.

6. Context-Aware Ranking

In this section, we introduce our proposed context-aware ranking algorithm. Here context is constituted by a fact and its surrounding facts. To be specific, for fact keyword set K_f ={lemon, vitamin c, rich}, its hypernym K_s such as {citrus fruit, vitamin c, rich}, its hypernym's hypernym K_s such as {citrus fruit, vitamin, rich}, its sibling K_s such as {grapefruit, vitamin c, rich}, {lemon, b complex vitamin, rich}, together with its hyponym, its hyponym's hyponym compose its context.

6.1 Hierarchical graph construction based on dictionary information

Since we consider one of relationships between two facts is generalization · specialization, we prefer to use this structure to rank fact keyword sets. The idea is to construct a hierarchical graph in accordance with hypernym/hyponym information from dictionary to include all keyword sets.

Given a query fact, as we mentioned in Section 5., firstly it is divided into words by taking morphological analysis so that a keyword set is obtained. Still take fact *"Lemon is rich in Vitamin C"* for example. At the very beginning, its keyword set K_f ={lemon, vitamin c, rich} is obtained. Based on substitution of only one element in K_f at a time, we can construct a subgraph shown as the enclosed area in Figure 3. To be more specific, since lemon is a kind of citrus fruits, in other words, "citrus fruit" is a direct hypernym of "lemon", the substitution of "lemon" to "citrus fruit" leads to get a parent keyword set {citrus fruit, vitamin c, rich} of K_f. So we add an edge between these two keyword sets. Similarly, "vitamin" is a direct hypernym of "vitamin c", the substation of "vitamin c" to "vitamin" leads to get a parent keyword set {lemon, vitamin, rich} of K_f. And an edge is also added between these two keyword sets. Pay attention that we assume that two keyword sets are connected by an edge if and only if the edit distance [注8] between them is 1. Following information in the dictionary, i.e. direct hypernym, direct hyponym, sister term, the nearby surrounding of K_f can be linked. Other keyword sets are then inserted into the graph relatively. For example, for keyword set {citrus fruit, vitamin, rich}, one of its elements "vitamin" is a direct hypernym of "vitamin c", so we consider this set as a parent keyword set of {citrus fruit, vitamin c, rich} and connect them by an edge. On the other hand, for each keyword set obtained from web search, if it shares any common parent set with any existing keyword set in the present graph, add itself and all keyword sets edited to the common parent set; if it shares no common parent set with any existing keyword set in the present graph, eliminate it. For example, by searching the web through the query "(*) ∧ vitamin c ∧ rich", a candidate keyword set {acerola, vitamin c, rich} is achieved. By looking up the dictionary, we know that a direct hypernym of "acerola" is "berry", while a direct hypernym of "berry" is "edible fruit". In the meantime, we find that a direct hypernym of "citrus fruit" is "edible fruit", which means that "lemon" and "acerola" share the same ancestor "edible fruit". Therefore, the keyword set {acerola, vitamin c, rich} and all keyword sets (here {berry, vitamin c, rich}, {berry, vitamin, rich}) edited to the common parent set {edible fruit, vitamin, rich} are inserted into the graph. Meanwhile, use dictionary to extend these new inserted keyword sets.

6.2 Cognition calculation for keyword sets

In short, our basic idea is to rank keyword sets based on the difference with its context. As a result, here we introduce the benchmarks to compare keyword sets. We use people's cognition of a fact to indicate how this fact is known by people. Besides, we assume that if a fact appears frequently on the web, it is well known by most people. Otherwise, it is unknown by most people. Hence, we take a frequency-based way to calculate cognition of the keyword set of a fact.

For the keyword set of a fact $K_f = \{k_1, k_2, ..., k_n\}$, how well it is cognized by people is computed as follows:

$$Cog(K_f) = \frac{{}_nC_{n-1}}{\sum_{q \in Q_d} \frac{HitCount(k_1 \wedge k_2 \wedge ... \wedge k_n)}{HitCount(q)}}$$

where $HitCount(q)$ indicates the hitcount returned by search engines, such as Google [注9], Bing [注10], when q is queried. Q_d represents the set of all kinds of combinations of any $n-1$ elements from K_f. n is the number of elements in K_f.

Take K_f ={lemon, vitamin c, rich} for example. How well

(注7)：Asterisk part corresponds to an entity in the extracted sentence.

(注8)：The edit distance between two keyword sets is the minimum number of single-keyword substitutions required to change one keyword set into another.

(注9)：https://www.google.com

(注10)：http://www.bing.com

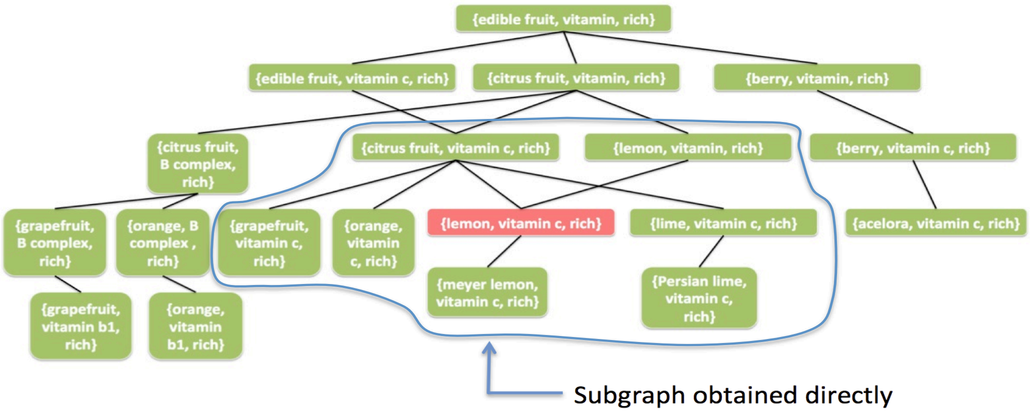
{edible fruit, vitamin, rich}
{edible fruit, vitamin c, rich}
{citrus fruit, vitamin, rich}
{berry, vitamin, rich}
{citrus fruit, B complex, rich}
{citrus fruit, vitamin c, rich}
{lemon, vitamin, rich}
{berry, vitamin c, rich}
{grapefruit, B complex, rich}
{orange, B complex , rich}
{grapefruit, vitamin c, rich}
{orange, vitamin c, rich}
{lemon, vitamin c, rich}
{lime, vitamin c, rich}
{acelora, vitamin c, rich}
{grapefruit, vitamin b1, rich}
{orange, vitamin b1, rich}
{meyer lemon, vitamin c, rich}
{Persian lime, vitamin c, rich}
Subgraph obtained directly

Figure 3 An example for part of constructed hierarchical graph for fact "Lemon is rich in Vitamin C". Each node indicates a keyword set.

it is cognized by people is determined by the following three parts:

（1） When people talk about things which is rich in Vitamin C, to what extent they can think up "lemon", say $\frac{HitCount(lemon,vitaminc,rich)}{HitCount(vitaminc,rich)}$.

（2） When people talk about which kind of thing lemon is rich in, to what extent they can think up "vitamin c", say $\frac{HitCount(lemon,vitaminc,rich)}{HitCount(lemon,rich)}$.

（3） When people talk about the relation between lemon and Vitamin C, to what extent they can think up "rich", say $\frac{HitCount(lemon,vitaminc,rich)}{HitCount(lemon,vitaminc)}$. Actually, the cognition for K_f ={lemon, vitamin c, rich} is the harmonic mean of the above three parts.

However, if the substitution k_1' of k_1 is not as widely discussed as k_1, it will lead a small hitcount of querying "$k_1' \wedge k_2 \wedge ... \wedge k_n$". Thus, the above calculation for cognition biases against those not so widely discussed. To solve this problem, we also take the cognition of each substitution into account, then

$$Cog(K_f) = \frac{2^n - 2}{\sum_{q \in Q_a} \frac{HitCount(k_1 \wedge k_2 \wedge ... \wedge k_n)}{HitCount(q)}}$$

where Q_a represent the set of all kinds of combination of elements from K_f, except K_f itself and $\emptyset$.

Therefore, for K_f ={lemon, vitamin c, rich}, besides the above three parts, its cognition is also determined by how well people cognize every element in K_f, say $\frac{HitCount(lemon,vitaminc,rich)}{HitCount(lemon)}$, $\frac{HitCount(lemon,vitaminc,rich)}{HitCount(vitaminc)}$, $\frac{HitCount(lemon,vitaminc,rich)}{HitCount(rich)}$, respectively.

There appears another problem that when people talk about the term "rich", they hardly mention about lemon and Vitamin C. Such kind of rare connection is likely to bring an unrealistic calculation result of cognition for K_f ={lemon, vitamin c, rich}. So we eliminate such part(s) in the process of cognition calculation.

6.3 Importance calculation based on context cognition

We briefly mentioned that we rank keyword sets based on the difference with its context. Actually, according to the hierarchical structure of the keyword set graph, we consider the importance of a keyword set is determined by (1) influence from its parent sets (only the direct ones); (2) influence from its sibling sets (ones which shares the same direct parent).

6.3.1 Influence from its parent sets

To simplify, let us concentrate on the blue node in Figure 4 which indicates a keyword set. Suppose there are two parent sets. In the left side, we can see one of its parent set P_1 is well known by people, marked high cognition in the figure, another P_2 is also well known by people. The blue node itself is highly cognized, too. Since the blue node is the child of P_1 and P_2, it is normal that the blue one inherits some attributes from its parents. Hence, in this case, that the blue node is highly cognized is to be expected. In other words, because the difference between its cognition and its parents's cognition is small, its parents are important rather than itself. We believe the influence from its parent sets is small. In the right side, we can see both of its parent sets P_1 and P_2 are unknown by people, marked low cognition in the figure. However, the blue node is well known. In this case, from parent to itself, there occurs a change which is beyond one's imagination so that the blue node is more important compared to its parents. We believe the influence from its parent sets is great.

To summarize, we assume the bigger the cognition difference between one and one's parent sets is, the more impor-

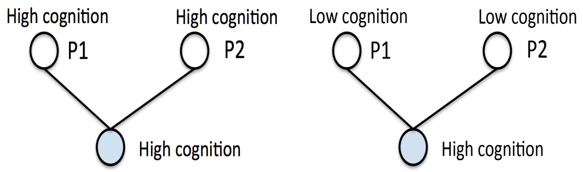

Figure 4 Illustrations of how one's parent sets influence one's importance.

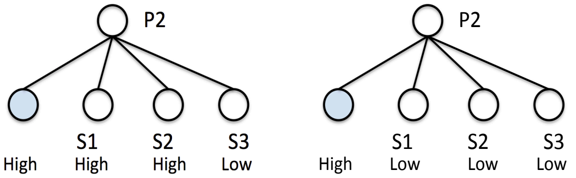

Figure 5 Illustrations of how one's sibling sets influence one's importance.

tance it is. Thus, influence from one's parent sets is calculated as:

$$ParentInf(K_f) = \sum_{p \in P} |Cog(p) - Cog(K_f)|$$

where P indicates the set of K_f's parent sets.

6. 3. 2 Influence from its sibling sets

Similarly, to simplify, let us concentrate on the blue node in Figure 5. Suppose there are three sibling sets. In the left side, we can see that among four child sets of P_2, there are three sets which is highly cognized by people, only one set which is unknown by people. Also, from the left side of the figure, we know that the blue node itself is well known. That means our concentrated blue node follows the general trend, and it is only a common one among the majority. Hence, it is not so important. In the right side, we can see that among four child sets of P_2, there are three sets which is unknown by people, only one set which is well known by people. Besides, from the right side of the figure, we know that the blue node is the one which is highly cognized by people. That means our concentrated blue node is an exception so that it should be presented to users. Hence, it is an important keyword set.

To summarize, we assume the bigger the cognition difference between one and the majority is, the more importance it is. Thus, influence from one's sibling sets is calculated as:

$$SiblingInf(K_f) = |\frac{1}{|M|} \sum_{m \in M} Cog(m) - Cog(K_f)|$$

where M is composed by sets which belongs to the major cognition part.

From the above discussion, we can draw the conclusion that context constituted by the input fact and its surrounding should be taken into account in order to bring more meaningful and important facts in the upper to users. Therefore, the final importance score for K_f is

$$Score(K_f) = \alpha ParentInf(K_f) + \beta SiblingInf(K_f)$$

where α and β are weight factors.

7. Conclusion

In this paper, we propose a system that given a fact, surrounding information of the fact are discovered, and then ranked according to cognition difference with context formed by surrounding facts.

Acknowledgment

This work was supported in part by the following projects: Grants-in-Aid for Scientific Research (Nos. 24240013, 24680008) from MEXT of Japan.

References

[1] catherine De Marneffe, M., Rafferty, A.N., Manning, C.D.: Finding contradictions in text. In: Proceedings of ACL2008:HLT (2008)

[2] R., R.D., Hongyan, J., Malgorzata, S., Daniel, T.: Centroid-based summarization of multiple documents. Inf. Process. Manage. 40(6), 919–938 (Nov 2004)

[3] Sanda, H., Andrew, H., Finley, L.: Negation, contrast and contradiction in text processing. In: Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1. AAAI'06, vol. 1, pp. 755–762. AAAI Press (2006)

[4] Tanaka, K., Yoshikawa, M.: Towards abstracting complex database objects: Generalization, reduction and unification of set-type objects. In: ICDT '88. pp. 252–266 (1988)

[5] Yamamoto, Y., Tanaka, K.: Finding comparative facts and aspects for judging the credibility of uncertain facts. In: Web Information Systems Engineering - WISE 2009, WISE2009, vol. 5802, pp. 291–305. Springer Berlin Heidelberg (2009)

[6] Yamamoto, Y., Tanaka, K.: Towards web search by sentence queries: Asking the web for query substitutions. In: Proceedings of the 16th International Conference on Database Systems for Advanced Application (DASFAA 2011). pp. 83–92 (2011)

[7] Yamamoto, Y., Tezuka, T., Jatowt, A., Tanaka, K.: Honto? search: Estimating trustworthiness of web information by search results aggregation and temporal analysis. In: Proceedings of the Joint 9th Asia-Pacific Web and 8th International Conference on Web-age Information Management Conference on Advances in Data and Web Management. APWeb/WAIM'07, vol. 4505, pp. 253–264. Springer-Verlag, Berlin, Heidelberg (2007)

[8] Yamamoto, Y., Tezuka, T., Jatowt, A., Tanaka, K.: Supporting judgement of fact trustworthiness by considering temporal and sentimental aspects. In: Web Information Systems Engineering - WISE 2008, WISE2008, vol. 5175, pp. 206–220. Springer Berlin Heidelberg (2008)