Proceedings of KFIS 2001 Fall Conference, 2001. 12. 1

망각에 의한 기억

Memorization by Oblivion

이중우 · 손세호 · 권순학
Jung-Woo Lee, Seo-Ho Son, Soon-Hak Kwon
영남대학교 전기전자공학부

요 약

본 논문은 무한 지식베이스에 가까운 웹으로부터 추출된 지식의 최적화 관리에 관한 것이다. 비록 웹문서로부터 사실이나 규칙과 같은 유용한 지식을 추출했다 하더라도 일반화되지 않은 지식을 포함하고 있으므로 이를 적절하게 제거함으로서 지식베이스가 일반화된 지식만을 포함하도록 관리해야 할 필요가 있다. 이를 위하여 본 논문에서는 인간의 망각에 기반한 기억방식을 응용한 망각에 의한 기억알고리즘을 제안한다. 본 논문에서는 기억을 관심도, 망각정도와 시간의 함수로 가정한다. 즉, 관심 있는 지식을 더 잘 기억하고, 잘 망각할수록 그리고 기억된 지 오래될수록 기억은 지수함수 적으로 감소한다. 여기서, 망각이란 이전의 기억정도, 기억능력 그리고 자극횟수의 함수로서, 이전에 기억된 정도가 크고, 기억능력이 크고, 자주 자극 받을수록 그 지식은 덜 망각하게 된다.

ABSTRACT

This paper is for the optimized management of the knowledge abstracted from the World-Wide Web(WWW) in which we assume the infinite knowledge-base. Though we can abstract various useful knowledge such as the facts and the rules from the WWW pages, they may include many noisy knowledge. Therefore we have to reasonably reject them from the knowledge-base which is composed of knowledge abstracted from the WWW. To do this, we propose the oblivious memorization concept. This concept is characterized by the memorization based on the oblivion mechanism of human being. We assume the memorization is the function of the concern for any knowledge, oblivion ability and time. That is, the more concern for any knowledge the more memorizable. And, the more oblivious and the more time spent the less memorizable by exponentially. Where, we assume the oblivion is the function of the degree of previous memorization, memorization ability and the number of knowledge stimulation. That is, the more previously memorized, the greater memorizing ability and the more frequently stimulated by any knowledge the less knowledge oblivious.

Keyword : Knowledge Base, Oblivion, Oblivious Memorization, Database Management

I. 서 론

월드 와이드 웹 (웹)의 놀라운 발전으로 하이퍼 링크를 찾아다니는 방법에 의해서 엄청나게 많은 웹 문서를 컴퓨터를 통하여 입수 할 수 있게 되었다. 이미 웹은 세상에서 가장 방대하

고 다양한 정보원이 되었고, 많은 전문가들은 앞으로 수십 년에 걸쳐 가장 기초적인 지식원으로 성장할 것으로 기대하고 있다.

비록 엄청난 양의 웹 문서를 입수하는 것이 가능해 졌지만 여전히 컴퓨터는 이러한 웹 문서의 내용을 인식할 수 없다. 만약 이러한 방대한 정보원인 웹 문서로부터 지식정보를 추출하고 이를 컴퓨터가 이해할 수 있는 기호나 통계적인 형태로 저장하여 지식베이스를 구축할 수만 있다면 이것은 인공지능 연구자뿐만 아니라 인터넷을 통한 정보검색의 분야에서도 획기적인 발전을 가져올 것이다.

그렇지만, 우리가 웹 문서로부터 관심 있는 주제에 관한 지식정보를 다량 추출했다 할지라도 아직 고려해야할 몇 가지 문제가 남아있다.

예로서 사과의 색상에 대한 지식정보를 추출하는 경우를 생각해 보자. 다음과 같은 문장을 웹 문서로부터 얻었다고 가정해 보자.

"사과는 빨간 색을 띠고 있다. 그러나 잘 익기 전에는 사과의 색은 초록색이다. 검정색 사과가 열린다면 재미있을 것이다?"

물론 자연언어를 정확하게 해석할 수 있는 기술이 있다면 사과는 빨간색이며, 조건부로 초록색을 띨 수 있고, 검정색이 아니라는 사실을 알 수 있을 것이다. 그러나 현재로서는 이와 같이 정확한 인식은 매우 어려운 상태이다. 그래서 현재의 자연어 인식기술로 아래와 같은 세 개의 지식정보를 얻었다고 가정해 보자.

색상(사과,빨강),
색상(사과,초록),
색상(사과,검정).

물론, 각각의 지식정보에 확신도를 매긴다면 서로 다른 값을 가지겠지만 상식적으로 우리는 세 번째 지식정보는 옳지 않고, 앞의 두 개는 어느 정도 옳다는 것을 알고 있다. 전문가 시스템의 지식베이스에 포함된 지식정보의 불확실성을 표현하는 전통적인 방법은 영에서 일의 값을 부과하는 확률적인 방법이다.

그러므로 위의 세 개의 지식정보를 확률적 방법에 기반한 확신도를 매긴다면 다음과 같이 될 수 있을 것이다.

```
belief( 색상(사과,빨강),
        색상(사과,초록),
        색상(사과,검정))
= (0.70, 0.28, 0.02).
```

그런데, 만일 사과가 빨갛다는 지식정보가 새로운 웹 문서의 인식결과 추가되었다면 어떤 일이 일어나겠는가? 물론, 색상(사과, 빨강)에 대한 확신도는 높아지는 반면 나머지 두 가지 지식정보는 그 확신도가 어느 정도 낮아 질 것이다. 인간의 기억하는 메커니즘에 비추어 볼 때 이것이 타당한 것인가? 사과의 색상은 그 종류에 따라 빨간색 일수도 초록색 일수도 있다. 즉, 여러 개의 지식정보가 서로 공존할 수도 있는 것이다. 어느 한가지 지식정보에 대한 확신도가 또 다른 사실의 확신도에 영향을 미칠 경우 많은 문제를 야기한다. 위의 예와 같은 경우 사과의 색상이 빨강과 초록이 모두 옳다고 가정하면 두 명제의 확신도는 결코 0.5를 넘지 못한다. 결국 추론시스템은 사과가 빨간색이라는 명제를 확실하지 않은 사실로 인식하게 된다.

이러한 문제를 해결하기 위하여 본 논문에서는 인간의 기억 메커니즘에 기반한 새로운 기억 알고리즘을 제안한다. 제안된 알고리즘에서는 기억을 얻고자 하는 지식정보에 대한 관심도와 망각정도와 시간의 함수로 가정한다. 즉, 관심 있는 지식을 더 잘 기억하고, 잘 망각할수록 그리고 기억된 지 오래될수록 기억은 지수함수 적으로 감소한다. 여기서, 망각이란 이전의 기억정도, 기억능력 그리고 자극횟수의 함수로서, 이전에 기억된 정도가 크고, 기억능력이 크고, 자주 자극 받을수록 그 지식은 덜 망각하게 된다.

II. 망각적 기억의 모델링

기억도, 관심도, 망각도, 기억능력, 지적자극과 같은 망각적 기억 알고리즘의 실현을 위한 몇 가지 개념을 정의한다.

[정의 1] 기억도, $m \in [0,1]$.

기억도란 두뇌와 같은 저장장치에 저장된 지식정보의 회상 가능정도로 정의한다.

이것은 인공지능 분야에서 자주 사용되는 지식정보에 대한 확신도의 개념으로도 사용될 수 있다. 즉, 높은 기억도(m)를 가지는 지식정보는 일반화된 지식으로 간주 할 수 있는데, 앞의 예의 경우 잘 구성된 지식베이스의 경우 색상(사과,빨강)은 높은 기억도를 가지는 반면, 색상(사과,검정)은 낮은 기억도를 가지게 된다.

[정의 2] 관심도, $\eta \in [0,1]$.

관심도란 주어진 시점에서 시스템이 주어진 지식정보에 대하여 관심을 가지는 정도로 정의한다.

인간은 자신이 관심을 가진 분야의 지식정보에 대하여 더 잘 기억하는 경향이 있다. 이를 이용하면 지식베이스에 시스템이 관심 있는 지식을 위주로 지식정보를 저장하게 할 수 있다.

[정의 3] 망각도, $\alpha \in [0,\infty)$.

망각도란 기억된 지식정보의 회상을 방해하는 정도로 정의한다.

망각하는 능력은 인간이 중요하거나 관심 있는 분야의 지식을 집중적으로 기억하기 위한 좋은 능력이다. 만일 인간이 한번 기억한 것을 영원히 명확하게 기억한다면 주된 지식정보의 기억에는 오히려 방해가 될 것이다.

망각하는 능력을 지식베이스의 관리 알고리즘에 추가함으로써 색상(사과,검정)과 같은 잡음에 해당하는 지식정보는 망각하게 되고, 결과적으로 지식베이스에는 일반화된 지식정보만이 남게 되어 확신도가 높은 지식으로 구성된 좋은 지식베이스를 얻을 수 있고, 이러한 지식베이스는 추론에 필요한 지나치게 많은 명제가 되는 지식정보로 인한 시스템의 복잡성을 감소시킴으로서 그 성능향상에 도움이 된다.

[정의 4] 기억능력, $q \in (0,\infty)$.

기억능력은 입력된 지식을 잘 회상하는 능력으로 정의한다.

인간은 서로 다른 기억능력을 가지고 있다. 어떤 사람은 한번들은 것을 잘 기억해 내는 반면, 어떤 사람은 여러 번들은 것을 잘 기억해 내지 못하는 경우도 있다. 그러나 전자가 반드시 후자보다 더욱 지능적이라고 말할 수는 없을 것이다. 왜냐하면, 지능이란 저장된 지식정보의 양 뿐만 아니라 그 지식정보로부터 추론에 의하여 다른 지식을 만들어 내는 능력을 포함하기 때문이다. 그러므로, 잡음에 해당하는 지식정보를 포함한 지나치게 많은 지식정보가 지식베이스에 저장될 경우 추론단계에서 너무 많은 대가를 지불해야 할 뿐만 아니라 잘못된 추론결과를 가져 올 수도 있으므로, 지식베이스에는 적절한 수준의 일반성이 있는 지식정보를 유지하는 것이 중요하다.

[정의 5] 지적자극

지적자극이란 시스템이 외부로부터 지식정보를 얻는 것을 말하며, 지적자극을 받는 횟수는 동일한 지식정보가 시스템에 주어지는 횟수로 정의하며, n으로 나타낸다.

인간은 동일한 지적자극을 자주 받을수록 잘 회상하는 경향이 있으므로, 지적자극횟수가 많을수록 망각도는 떨어지게 된다.

본 논문에서는 망각도를 아래와 같이 모델링 한다.

$$\alpha = \frac{(1 - m_{t-1}^{1/q})^q}{n^2} \qquad (1)$$

여기서, m_{t-1}은 이전시점에서의 기억도.

그러면, 현시점에서의 기억도는

$$m_t = \eta \cdot \exp(-\alpha^2 \cdot \Delta t^2) \qquad (2)$$

와 같이 모델링 될 수 있다.

여기서, Δt는 직전 지적자극 후 현 지적자극까지 소요된 시간이다.

III. 모의실험 및 결과

제안된 망각에 의한 기억 알고리즘은 웹 문서로부터 추출된 지적정보를 지식베이스에 저장하는 단계에서 지식베이스에 잡음에 해당하는 일반적이지 못한 지적정보를 제거함으로서 일반화된 양질의 지식정보만이 남도록 '함으로써 지식베이스의 질적 향상을 도모하는 것과 같은 다양한 분야에 응용될 수 있다.

본 논문에서는 지식베이스에 저장된 지적정보의 기억도를 그 지적정보의 확신도로 사용함으로서 일반화된 양질의 지식베이스를 유지하도록 하는데 제안된 알고리즘을 응용하고자 한다. 그림 1.은 제안된 응용에 대한 개념도 이다.

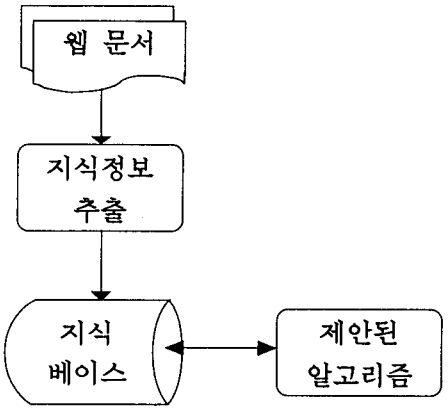

그림 1. 제안된 알고리즘의 지식베이스 관리에의 응용 개념도.

Fig. 1. Conceptional diagram of application to the management of knowledge base by the proposed algorithm.

초기에 주어진 웹 문서로부터 하이퍼텍스트 링크를 따라가면 많은 웹 문서들을 얻게 된다. 이렇게 얻어진 웹 문서들은 HTML(Hyper Text Markup Language)형식이나 순수한 텍스트형식이나 기타 문서를 표현하는 다양한 형식을 가지는 문서이다. 이러한 문서들은 자연어처리에 의해서 지식정보를 추출하기 위해서는 순수한 텍스트형식의 문장으로 변경해야 한다.

발표된 논문들이 이 단계에서 일반성을 가지는 양질의 지식정보를 추출하는데 초점을 맞추고 있으나, 현재의 자연언어처리 기술로는 일반성의 관점에서 지식정보의 질적 향상을 도모하는 것은 상당한 비용을 필요로 한다. 그러므로, 추출된 지식정보의 일반성을 매기는 기준으로 확률적 관점에서의 확신도(Belief)를 사용해 왔으나 이 또한 여러 가지 문제점을 포함하고 있다. 몇 가지를 살펴보면,

표 1. 확률기반 확신도와 제안된 알고리즘에 의한 확신도의 비교.

Table 1. Comparison of belief value between probability-based and the proposed.

추출된 시점 (월/일/시)	추출된 지식정보	확률에 의한 확신도	제안된 알고리즘에 의한 확신도	
			$\eta=0.5, q=2.0$	$\eta=0.5, q=2.5$
01/01/00	색상(사과, 빨강) 색상(사과, 초록)	{0.50, 0.50}	{0.50, 0.50}	{0.50, 0.50}
01/02/00	색상(사과, 빨강) 색상(사과, 검정)	{0.50, 0.25, 0.25}	{0.50, 0.01, 0.50}	{0.65, 0.31, 0.50}
01/03/00	색상(사과, 빨강)	{0.60 ,0.20, 0.20}	{0.67, 0.00, 0.01}	{0.83, 0.07, 0.31}
01/04/00	색상(사과, 초록)	{0.50, 0.33, 0.17}	{0.69, 0.50, 0.00}	{0.83, 0.50, 0.07}
01/05/00	색상(사과, 빨강)	{0.57, 0.29, 0.14}	{0.84, 0.39, 0.39}	{0.91, 0.49, 0.01}
01/07/00	없음	{0.57, 0.29, 0.14}	{0.84, 0.05, 0.00}	{0.91, 0.39, 0.00}
01/10/00	없음	{0.57, 0.29, 0.14}	{0.84, 0.00, 0.00}	{0.91, 0.18, 0.00}

1. 잡음으로 간주되는 지식정보의 증가는 우리가 원하는 일반성을 가지는 양질의 지식정보의 확신도를 감소시킨다.

예를 들어 추출된 지식정보를 k_i로 표시할 때 추출된 지식정보가 $\{k_1, k_1, k_2\}$라면, k_1의 확률적 확신도는 2/3=0.66인 반면, $\{k_1, k_1, k_2, k_3, k_4, k_5\}$의 경우 2/6=0.33으로 감소되어 버린다.

2. 추출되는 지식정보의 양이 늘어남에 따라 지식베이스의 용량이 계속 증가한다.

예를 들어 위의 1.의 예의 경우 원하는 지식정보는 k_1으로 한 개인 반면 전자의 경우 세 개를 후자의 경우 여섯 개의 지식정보가 모두 저장하고 있어야만 확률적 확신도를 계산 할 수 있다. 왜냐하면, "각각의 확신도의 합은 1이다"라는 확률적 개념 때문에 비록 잡음으로 판단되는 지식정보라 할지라도 지식베이스에서 삭제시킬 수가 없다. 그렇지만, 제안된 망각에 의한 기억 알고리즘을 적용할 경우 기억도에 의해 표현되는 확신도가 매우 낮을 경우 해당하는 지식정보를 지식베이스에서 삭제하여도 연관된 다른 지식정보의 확신도에는 영향을 미치지 않는다.

표 1.은 웹 문서로부터 추출된 지식정보의 입력들로부터 확률적 확신도를 계산한 결과와 본 논문에서 제안된 알고리즘을 응용한 확신도를 계산한 결과를 비교하고 있다. 일반성을 가지는 지식정보는 잡음으로 볼 수 있는 지식정보 보다는 빈번하게 웹 문서에 나타난다고 볼 수 있으므로 위의 결과에서 색상(사과,빨강)이 오랫동안 확신도를 높은 값으로 유지한 반면, 잡음으로 볼 수 있는 색상(사과,검정)은 시간이 지나면서 점점 망각되어져 낮은 확신도를 가지게 된다.

3. 모의실험 및 결과

본 논문에서는 망각에 의한 기억 알고리즘에 대하여 소개하고 있다. 그리고, 지식베이스가 일반성을 가지는 지식정보들을 포함하는 양질의 지식베이스가 되도록 관리하기 위하여 제안된 알고리즘에 의한 확신도를 계산함으로서 그 유용성을 보였다.

앞으로의 연구과제는 실제로 대량의 웹 문서에서 추출된 지식정보들을 포함하는 지식베이스에의 적용을 통하여 실제적인 유용성을 입증하는 것이다.

IV. 참고문헌

[1] R. Fagin and J. Y. Halpern, "Uncertainty, belief, and probability," Computational Intelligence, vol. 7, no. 3, pp. 160-173, 1991.

[2] E. Riloff and W. Lehnert, "Information Extraction as a Basis for High-Precision Text Classification," ACM Transactions on Information Systems, vol. 12, no. 3, pp. 296-333, 1994.

[3] D. Freitag, "Information extraction from html: Application of a general learning approach," in Proc. Fifteenth Conf. on Artificial Intelligence AAAI-98, pp. 517-523, 1998.