语义网中基于用户意图的关联资源搜索和排序算法

Searching and Ranking method of relevant resources by user intention on the Semantic Web

汇报人：王琦

目录页

CONTENTS PAGE

研究议题介绍

文献综述

研究过程

讨论和建议

研究议题介绍

研究背景
研究动机
研究意义

随着internet信息的显著增长，信息检索过程中遇到了更多的限制，为了适用于用户的使用，网页设计过程中更多的是图文混合。为了克服这些缺陷，w3c提出了语义网（基于本体）来促进web检索，为了达到语义检索的目的，语义网必须提供基于资源间不同关联的检索方法。
本文中，作者提出了一种予以关联检索方法（包括资源评价和资源间的关系，同时还有基于本体和资源属性语义网的相关信息识别）。提出的语义检索方法主要基于扩展激活扩散技术。为了评价搜索结果的重要程度，研究者提出了一种权衡方法基于搜索结果的特性和一般性来衡量属性和资源。通过这个研究，用户款可以检索到对他们更有价值更重要的信息，实验结果证明该方法在检索和排列语义检索结果上是有效并且高效的。

Introduction

多说信息检索系统通过计算被检索对象和用户检索条件的匹配数值来对检索结果进行排序，然后选择最前面的对象返回给用户。第一代自动信息检索系统产生于50-60年代，这些系统都要强调信息检索技术。自那以后，蒂姆伯纳斯李提出一个超文本项目的建议，被称为world wide web，万维网于1989年开始，在网络技术和信息存储技术的推动下，使得数字化大文件成为可能。最后搜索引擎成为最适合进行大数据检索的途径。然而由于可用信息大规模增加，使得用户越来越难找到对自己有用的相关信息。一个重要问题时网页时文本和展现的混合形式，，超文本标记语言构成了网页的结构；另一个问题是搜索引擎主要基于关键词检索技术，一般的搜索引擎搜索通过收集，解析和存储网页关键词数据来促进精确信息检索。当用户键入搜索关键词后，搜索引擎通过索引和关键词去匹配搜索结果并进行排序。为了让用户能够得到更精确的信息，语义检索应该将语义独立于html语法，，搜索引擎也应该具备语义化搜索的能力，这中需求促使了语义网的诞生。

语义网是万维网的一次革命，使得web可以更好地满足用户的要求。为了达到这种目标，领域内的公共信息的词汇库（本题库）被定义，使用元数据和关联的方式对数据进行标记，本体包含了万维网上的资源、资源间的关联。当前web主要是通过网页和超链接形式实现的，然而本体提供了更复杂的网络结构，因为它包含概念以及概念的描述以及关系。我们把这个称为语义网。语义网应用可以获得更精确的搜索结果，，但是需要不同的检索方法来确定是否是用户需要的信息关联。语义检索和关键词检索一个最主要的不同就是对数据间相互关系的充分利用，在语义网中，这是一种资源。

Introduction

本文中提出的检索方法能够检索到与用户检索条件相关的信息和概念，即使一些相关信息并不存在外显的关联；这种检索能力主要基于分散激活方法，传统检索只要找到关键词即可，如输入关键词“Metaweb Technology”，然后搜索引擎就会定位包含关键词的文档并提供给用户查看。然而本文的方法基于扩散激活方法，确定与“metaweb technology”有语义关联的概念作为搜索结果。因此关联资源的属性在语义网中非常重要，因为他们解释了某个资源“为什么”“如何”与搜过的条件进行关联。传统网页通过超链接进行关联，而语义网则通过属性进行关联，不同的属性暗示了不同的关联权重；传统的搜索仅仅提供一个搜索列表，权重是按照specificity 和generality来分配的。但是本文中采用了一种非常有趣的权重计算方法。

本文中我们提供了一种基于扩散激活的语义搜索方法，并用来定位相关结果，他们与用户的检索都是语义相关的。这种方法能够检索到相关的所有的概念即便是关键词在文档中没有出现过。此外我们讨论了属性和资源的权重分配，以此来支持语义搜索，提供给用户合适的排序结果。换句话说权重分配过程和扩散过程都是经过检验的，最终我们通过真实环境下的检索对比来评价这种检索方法的效果和权重。
本文包括下面几个部分：相关研究，数据模型和权重方法，提出语义检索方法，语义检索系统的讨论和实验结果讨论，结论和展望。

文献综述

Part

2

Related Work

近年来很多语义检索方法被提出，他们应用领域多样，但他们都是基于一套公共的方法（Mongold，2007）。Manggold提出了一套将这些方法按照维度进行分类的分类方案（架构，耦合，透明性，用户上下文，查询修改，本体结构和本体技术），他选取了10种语义检索系统进行比较，通过研究得出了未来研究和应用发展的趋势，从他的研究角度看，我们的系统时高度耦合网页和本体的，这意味着文件源数据能够明确的揭示本体概念。因此我们的方法被分类为基于图的方法，能够感知本体论概念和文件作为图的节点。

The Multimedia E-Culture project(Schreiber et al,2008)是一个语义检索系统，揭示了如何部署语义网和呈现技术以从而为文化继承资源集合提供更好地索引和检索支持。为了检索语义路径，该系统会查看和遍历所有的RDF文本来匹配给定关键词，直到找到一个相关资源，最终，检索结果基于路径（与结果匹配的文本）进行聚类。这个研究与我们的研究有相似点，但是这个研究缺乏为属性和资源分配权重的能力和识别搜索结束阈的能力。这些限制是这个系统目前最重要的问题因为他们是扩展与语义网的决定因素。此外，该系统中的信息检索只有一条路径：即从查询三元组到相关主题。

一些语义网排序技术也被提出，SemRank基于可预测性进行排序，它以相关模型为基础，提供了从描述层面上衡量属性与其他相关属性相比的唯一性的方法。然而为了将扩散激活方法引入语义检索，不仅仅是描述，可用的资源也必须考虑，因此我们扩展了属性的很亮方法并提出了一种新方法来衡量资源的唯一性

Related Work

相关的检索中，信息间的关系往往被展现成一个网络，信息项是其中的一个节点，关联是节点之间的联系。激活扩散模型（Cohen&Kjeldsen， 1987）是一个用来检索和处理关联网络的方法，纯粹的扩散激活模型由网络数据结构的简单概念处理技术构成，检索过程首先进行初始化标注（权重，激活值），然后渐进通过冒泡和扩散的方法将激活渐进扩散到关联节点，这些权重值会随着冒泡而降低，。扩散的结果是节点的激活水平和语义路径通过不断遍历达到终点。然而纯粹的扩散激活模型有一些缺陷，因此提出了一些启发式推理规则来增强这种模型，距离限制，扇出约束，路径约束和激活约束通常用于激活扩散模型。 我们的研究就是基于增强的扩散激活模型，其中使用了一些约束作为终止条件。

Rss（Ning,Jin,&Wu,2008）一个语义检索排序框架，它正是基于扩散激活方法，该框架充分利用关系的异质性来很亮资源的重要性，例如，支持语义检索和为用户提供合适排序的搜索结果。这个系统人为的分配边缘权重并将其应用于实例水平。为了全局的对资源重要性进行排序，系统采用了随机检索模型，然后计算临界概率与PageRank标准概率一致的实体。然后采用扩展扩散激活散发检索语义关联最大的资源。然而，这种框架下，节点的权重并没有很好地反应扩散激活，因为他们都是初始化激活的值。为了解决这个问题，我们提出了一种解决方法来计算临界值和节点权重。

研究过程

Part

3

Calculating Edge and node weights

首先定义基于知识的语义网数据模型，然后讨论分配属性和资源权重的方法并规范化扩散激活。

3.1Definition of knowledge base

本体是代表领域知识的机制，包括概念等级，概念间的角色关系。语义网中应用本体来提供一种可理解的数据集，所有的信息都是语义关联的，所以用户可以搜索信息间的关系，因此，语义检索结果包含具有语义关联的概念。

我们的语义检索方法的目标是在获得扩散激活后对搜索结果进行排序，为了应用扩散激活方法和对结果进行排序，我们需要扩展本体的概念，从而使本体包含属性和实例的权重。一个扩展的知识单元可以被描述如下：

KB=（I,C,P），

C={c_i|c_i属于owl：class}，

P={<p_i,f(p_i)>|p_i属于rdf:Property,f(p_i)属于R^+},

I={<$i_i$mf(i_i)>|i_i属于c_j,f(i_i)属于R^+},

f代表衡量属性和实例权重的方法，f包含两个步骤，第一步是基于特异性和一般性计算权重，第二步是将权重的数值标准化，3.2会详细介绍。只有属性和实例有权重，因为搜索结果得出的是与搜索条件相关的实例的集合。

Owl有两种属性：对象属性显示了两个类中实例的关系，数据类型属性解释了类实例，RDF 文本和XML Schema数据类型之间的关系。我们的研究只扩展语义检索让其包含对象属性而不包括数据类型属性是因为数据类型不是实例间的关系，我们用数据类型来对比概念和检索语句之间的相似度并未概念提供额外信息。

Calculating Edge and node weights

3.2 Edge and nodeweights basedon specificity

信息论中，自信息用来衡量一件事情发生所包含的信息量的多少。概率事件中的自信息的量只取决于概率事件的本身，事件的概率越小，事件发生的可能性就越小，但是一旦发生其所包含的自信息量就越大。自信息的衡量有以下两个属性：如果事件C由两个独立的事件A和B组成，那么C的信息量就等于A和B的和，自信息则由事件发生概率的负对数来衡量，

S（$x=x_i$）=$\log(1/\log pr_i)=-\log pr_i$ pri是事件xi发生的概率

基于此，SemRank提出了很量自信息量的模型，通过考虑事件发生的临界值并把RDF属性作为输出。为了衡量自信息的临界值，他们定义了两个衡量参数：specificity和 θ-specificity. ，前一个属性是衡量与其他属性之间特异性的，后一个 是衡量相对于所有Abox中领域和范畴属于同一个语义网的属性的特异性。这两个参数共同决定了自信息的总量。

上述所说的SemRank提出的仅仅是衡量属性特异性的方法，然而为了在语义网中应用扩散激活方法，我们需要考虑节点的权重。相应的我们扩展了这种属性衡量方法，并提出了一种新的考虑了资源在网络扩展中特异性的方法

Pr(x=i)= $\Pr(\chi = i) = \frac{|(i,*,*) + (*,*,i) - (i,*,i)|}{|(*,*,*)|}$.

Pr(x=i)指的包含实例i的三元组的概率，他是实例i的特异性值，一个资源的特异性是衡量他uniqueness的重要参数，

$$S_S(i) = S(\chi = i) = -\log \Pr(\chi = i),$$

Calculating Edge and node weights

S(i)指的是实例I的自信息，他基于实例i相对于其他实例的出现概率
相应的，就有可能开发一个类似的利用语义RDD的衡量方法，在RDF Schema中，有两个属性，domain和range，用来描述属性和类是如何在在RDF数据中如何使用。Domain属性用来说明特定的属性可以应用于指定的对象，range属性用来书名特定属性的值是指定的类的实例。
如果一个属性存在domain或者range，

$$\Pr(\chi \in \theta_i) = \frac{|\theta_i|}{|(*,*,*)|} = \frac{|(*,p_i,*)|}{|(*,*,*)|}, \qquad \Pr(\chi = i|\chi \in \theta_i) = \frac{\Pr(\chi = i)}{\Pr(\chi \in \theta_i)} = \frac{|(i,*,*) + (*,*,i) - (i,*,i)|}{|(*,p_i,*)|}.$$

$$S_{\theta-S}(i) = S(\chi = i|\chi \in \theta_i) = -\log \Pr(\chi = i|\chi \in \theta_i).$$

In order to determine the total self-information of instances, we combine the self-information of *specificity* and *θ-specificity*:

$$S_i(i) = S_S(i) + S_{\theta-S}(i).$$

表示市里的节点在语义网中都有自己的权重（基于独特性），松耦合节点要比紧耦合的更有价值。节点的权重是用来进行语义网中的激活的和计算搜索结果排序的。

Calculating Edge and node weights

3.3Normalization of weight values

我们讨论了如何根据独特性给临界值和节点分配权重，然而，用上述方法计算出的特异值有庞大的数量，通常来讲这是因为属性的频次是数量较少的三元组，但是本体是巨大的，即时仅仅是一个专业领域的本体。因为我们用负对数来衡量权重值，如果边缘和节点的值接近于0，那么权重值将接近无穷，因而，这些权重值在扩散激活方法中不能直接拿来使用，我们定义了一种方法来标准不啊这些权重值为【0,1】。

为了把数列的值都汇聚到【0,1】，我们使用了自然对数。

$$P(t) = \frac{1}{1 + e^{-t}},$$

T值为负无穷时，p=0，t为正无穷时p=1，然而，我们衡量的权重值，基于负对数函数，值从0到正无穷，所以简单地自然对数不能够达到预期目的，所以我们使用了改进的自然对数

$$P(t) = \frac{1}{1 + e^{-y\cdot(t-x)}},$$

X指的是分布的均值，y指的是权重参数，换句话说，由x和y共同影响曲线，当T=x时，p=0.5，然后时期接近于标准正态分布，其标准差就是其中所有点与均值的偏离程度

$$\mathrm{NI}^{p}(p) = \frac{1}{1 + e^{-s_p\cdot(S_p(p)-\mu_p)}},$$

where $S_p(p)$ denotes the values of self-information of property p, μ_p denotes an average value of the edge weights $S_p(p)$, and s_p denotes a slope value of the logistic function for the normalized edge weights.

The normalized function for the weight of the nodes is defined by the formula:

$$\mathrm{NI}^{i}(i) = \frac{1}{1 + e^{-s_i\cdot(S_i(i)-\mu_i)}},$$

where $S_i(i)$ denotes the values of self-information of instance i, μ_i denotes an average value of the node weights $S_i(i)$, and s_i denotes a slope value of the logistic function for the normalized node

S(p)代表属性p自信息的值，up代表边缘权重的均值，sp代表在函数中的斜期中s(i)代表实例自信息得值，si代表斜率，、

现在我们讨论了边缘和节点权重的衡量方法，这个权重方法有助于找到松连接的信息。

Calculating Edge and node weights

3.4.Edge and node weight based on generality

这一部分，我们讨论了另一个衡量权重的方法，个性和一般性是两个相对的概念，与specificity不同，基于generality的方法中，edge和nodes中具有更多连接的往往会有更高的权重值。

为了很亮基于generality的权重值，我们定义了下面的公式 $G(\chi = x_i) = -\log(-pr_i + 1).$

当pr=0时，g=0，当pr接近1时，g接近于无穷。这在我们的研究中是一个问题，因为pr经常是1.比如说，如果只有一个类间属性被定义，那么他的概率就是1，在这种情况下我们分配了bigM作为它的权重（因为值是正无穷），计算generality权重的过程和计算specificity的类似，只有基本函数是不同的，其公式如下：

同样的为了将这些值标准化，我们也定义了NG方法

$$G_p(p) = \alpha \cdot \left(-\log\left(-\frac{|(*,p,*)|}{|(*,*,*)|}+1\right)\right) + (1-\alpha) \cdot \left(-\log\left(-\frac{|(*,p,*)|}{|(i_i,*,i_j)|}+1\right)\right),$$

where $i_i = \{instances\ of\ c | c \in p.domain\}, i_j = \{instances\ of\ c | c \in p.range\}.$

$$G_i(i) = \alpha \cdot \left(-\log\left(-\frac{|(i,*,*)+(*,*,i)-(i,*,i)|}{|(*,*,*)|}+1\right)\right) + (1-\alpha) \cdot \left(-\log\left(-\frac{|(i,*,*)+(*,*,i)-(i,*,i)|}{|(*,p_i,*)|}+1\right)\right),$$

Semantic Searching based on spreading activation

这一部分中我们呈现了我们的语义检索过程，4.1描述了基于增强扩散激活方法的语义检索过程，4.2描述了扩散激活模型的局限。

4.1.Semantic search algorithm

为了检索相关信息，我们使用了应用增强扩散激活算法的语义检索过程，这个算法在初始节点（与用户检索条件匹配的）放置一个specified初始激活值，然后程序会不断地迭代直到人工停止或者出发终止事件。迭代器包含一个扩散面和一个终止检测面，在扩散面中会不断地从初始节点向邻近的节点进行扩散。节点的激活权重也会在扩散函数中倍计算得出，在终止检测面中，当没有临接节点或限制条件满足时就会触发终止方法，扩散方法完成后就会获取到一个节点集，这些节点都是按照激活值进行排序。

在第一步中，初始节点时通过查询本体关键词定位的，初始集包括扩散过程的初始节点和每个节点的激活值。 Initial Set $IS = \{(i_1, w_1), (i_2, w_2), \cdots, (i_n, w_n)\}$,

理论上可以选择不同的权重初始激活值，但在本文中，全部选择初始激活值为1，即加入要检索“BLU”，那么在语义本体库中找到的初始节点就需要完全包含关键词“BLU”。然后会在初始集合的基础上扩散到关联的实例节点，当扩散到其他节点后他的输出值必须定义，实例j的输出值公式如下

$$A_j(t) = \tanh(IP_j(t)),$$
$$IP_j(t) = \underset{P_{ij} \in edges\ between\ i\ and\ j}{\mathrm{MAX}} (NI^p(p_{ij}) \cdot A_i(t)) + A_j(t-1),$$

Semantic Searching based on spreading activation

$$A_j(t) = \tanh(IP_j(t)),$$
$$IP_j(t) = \underset{P_{ij} \in edges\ between\ i\ and\ j}{\text{MAX}} (NI^p(p_{ij}) \cdot A_i(t)) + A_j(t-1),$$

其中t代表某个时间点，Aj(t)代表节点j在t时间的激活值，Ipj（t）代表t时间从i节点输入j节点的值，NI（pij）指的是链接i-j的标准化的属性p的权重。在时间0时每个节点的激活值都是0，节点的输出通过tanh（）来赋值。节点的输入值被定义为节点值乘以激活值乘以权重乘以该节点在前一时刻的激活值的最大值的和？

本文的主要目的时确定更独特和更一般的权重方法，使用这种方法，用户可以更好地考虑到边界和节点而不是只考虑边界权重。为了应用节点权重，我们扩展了输入函数

$$IP_j(t) = \underset{P_{ij} \in edges\ between\ i\ and\ j}{\text{MAX}} (NI^p(p_{ij}) \cdot NI^i(j) \cdot A_i(t)) + A_j(t-1),$$

NI（j）代表节点j的权重。

通过不断的扩散，就找到了查询结果，图1呈现了扩散激活算发的code，然而其中冒泡有可能会达到整个网络，为了解决这个问题，激活会加一些约束。

```
Set activationValue<node, value>
List paths
while(paths is not empty)
    path = remove(0) of paths
    preNode = getLastNode of path
    preActivationValue = getActivationValue of preNode
    Find liked nodes Nodes from preNode
    for each(node ∈ Nodes)
        if(visitedNode(node))
            exit
        max = Max(edgeWeight(from preNode to node) * preActivationValue)
        inputValue = max + getActivationValue of node
        activationValue = tanh(inputValue);
        Update activationValue of node
        Add node to path
        Add path to paths
```

Fig. 1. A pseudocode of the spreading activation algorithm.

Semantic Searching based on spreading activation

4.2.Constraints of spreading activation

为了防止扩散到整个网络，Cohen等人提出了一些约束，1.首先添加一个基于启发式规则的距离约束（在扩散过程中两个节点关联强度降低），当扩散距离超过初始设定的距离约束时扩散就应当停止。

2.扇出约束被应用于此来避免过度广度扩散。

3.路径约束可以使用权重和链接标记建立一个约束模型，如果链接标记，那么激活转向一个特定的路径，同时阻塞下面意义不大的路径。

4.激活约束有可能在单节点层面上通过使用阈值函数控制激活扩散，这种方法可以通过改变单个脉冲相对于整个网络的关系阈值来实现。

5.类约束可以在激活必须不能冒泡到指定类节点时使用。

在这些约束中，我们使用激活约束，距离约束和类约束，为了使用激活约束，我们设置了输入函数的阈值（0.1），从而使得激活不会在进行扩散。当阈值为0.1时只有输入边缘权重才会激活，这是被允许的因为我们假设这是不包括在90%标准化值中的离群值。设为0.01那么输入值会同时考虑node和edge权重，只有当输入值大于临界值时才会从当前节点冒泡到其他节点。

距离约束依据路径距离限制扩散，为了应用距离扩散，我们定义了激活函数：$A_j(t) = \tanh((1 - decayFactor \cdot depth) \cdot IP_j(t)),$ decayFactor指的是衰减变量，在冒泡过程中会减小，depth指的是从初始实例的冒泡距离，使用衰减变量的原因是实例间的管理强度随着距离的增加而降低，我们设置衰减参数为0.3，这样最大的冒泡升读限制为3层，Crestani认为设为3层已经足够。

Semantic Searching based on spreading activation

最后，类约束的目的是确定实例饱汉子指定的用户类中，如果用户设置了一个特定的类查询，那么扩散激活过程在于给定类类型的节点相交时就会进行检测和终止，一个用户可以使用这个约束简单地定位感兴趣的实例。

Implementation of a Semantic S

我们研究的目标是提出基于本体的语义检索方法，图2呈现了我们基于关联的检索系统的组成，网络数据通过语义标注转换成三元组结构被存储在语义元数据中，用户通过这个系统可以找到与他们查询有关的信息。

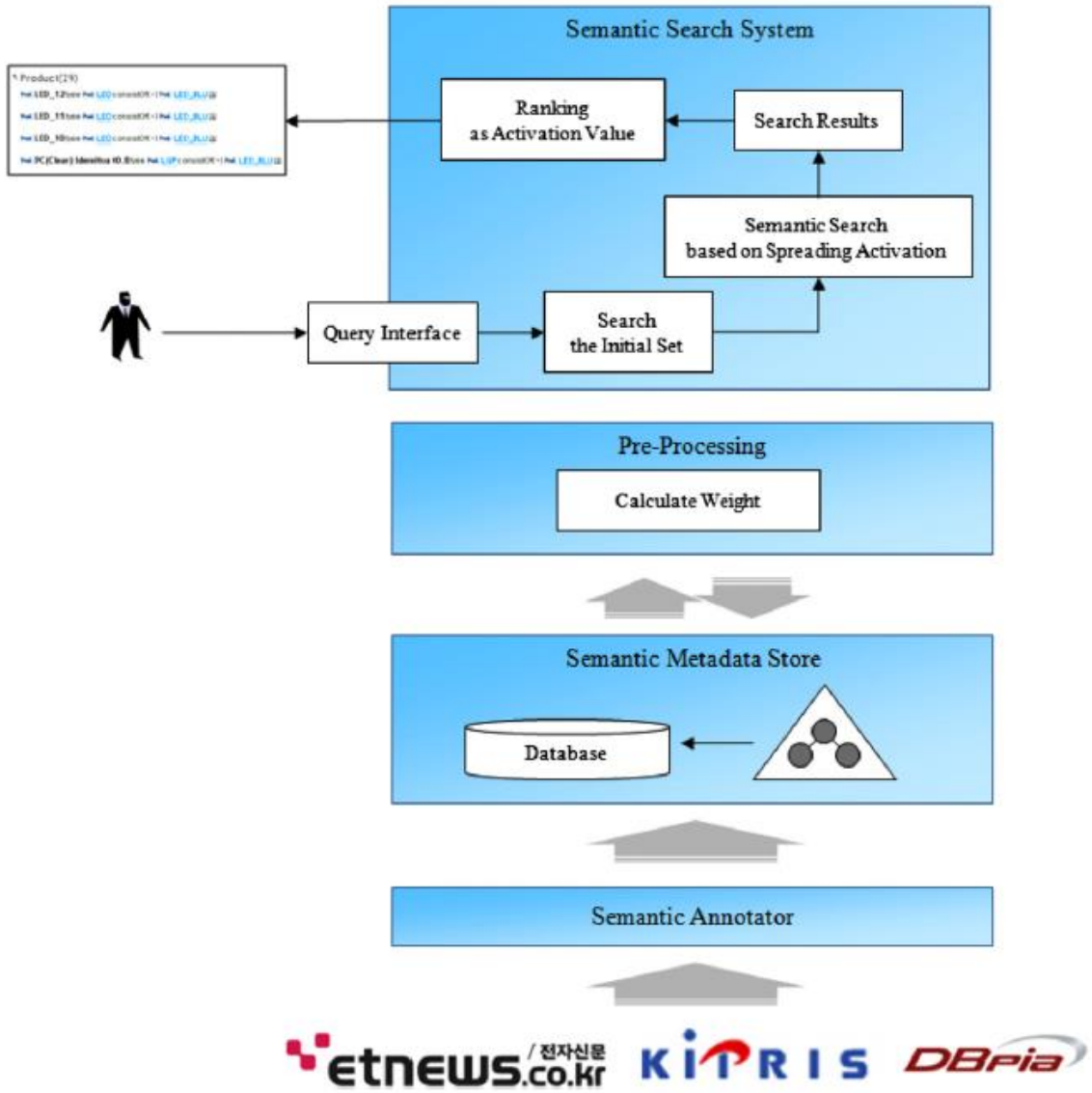

Fig. 2. Architecture of a semantic search system.

Implementation of a Semantic Search

网络数据被转化为基于本体的数据，本体被存储在本体数据库中，为了减少查询时间，激活过程中属性和实例的权重在与检索处理中被计算，因为这是一个耗时的过程。如果一个用户请求一个语义检索关键词，检所代理首先要确定开始检索的初始化集，扩散激活完成后，检索结果基于激活值进行排序，最后，用户就得到了相关的实例和资源。

为了检索语义信息，我们建立了一个领域本体，为了建立电子学方面的领域本体，我们搜集了电子学各个方面的信息，如新闻网站，专利服务网站和研究论文服务网站，同时使用了予以自动标注的方式，开始提取概念（主题，作者，关键词，url，出版日期，来源），然后建立关系三元组，我们从不同网站建立的本体包括七个方面的分类，（News，Person，Article，Product，Patent，Technology）但是数据源是不同的，这就意味着每个信息源都和其他信息源是分离的。虽然在独立信息源中可以检索，但在不同信息源之间建立信息关联是不可能的，这就导致了对新技术，趋势和零件生产商特别敏感的电子公司的员工会在读关于电子方面的文章时不得不在不同站点中搜索专利和技术相关的信息，本体检索系统恰好可以解决这个问题，由于这些方面的信息是语义关联的，例如一个专利关联了一个技术趋势单专利是被一个公司拥有的，那么，我们将原信息转化成本体这样他们就关联起来了，这是语义检索系统一个重要的改进。

图3是从一个新网站进行语义标注的结果，主要通过RDF和OWL实现，主要收集了上述七个方面的信息。

Implementation of a Semantic Search

11


```
<News rdf:ID="NW006">
    <dc_source rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
        http://www.etnews.co.kr/news/detail.html?id=200704230077
    </dc_source>
    <dc_title>
        <Title rdf:ID="Title_22">
            <koreanTitle rdf:datatype="&xsd;string">
                루미마이크로, 조명 및 플래시용 고휘도 LED 개발
            </koreanTitle>
        </Title>
    </dc_title>
    <dc_creator>
        <Reporter rdf:ID="Reporter_uhyeongjun">
            <name rdf:datatype="&xsd;string">유형준</name>
        </Reporter>
    </dc_creator>
    <dc_date rdf:datatype="&xsd;dateTime">
        2007-04-24T00:00:00
    </dc_date>
    <articleCode rdf:datatype="&xsd;string">
        NW006
    </articleCode>
</News>
```

Fig. 3. Semantic annotation from a news web site.

结果和结论

Part

4

Result

我们采用实验检验了我们的算法，为了验证系统的有效性，实验主要包括验证和评价，1.验证实验是确定搜索的正确性，2.评价主要是将我们的搜索结果与其他语义检索方法进行对比。
为了对比权重处理方法和扩散激活约束影响和效率，我们建立了一个简单地本体（如图4）

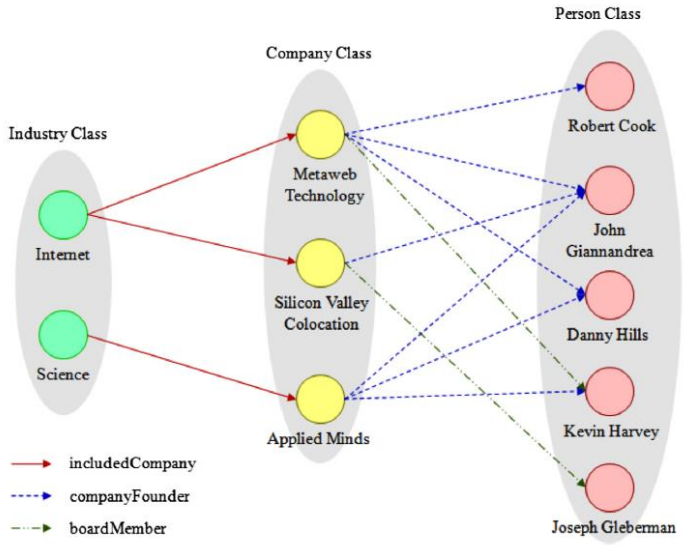

Fig. 4. Simple experimental ontology.

Result

我们使用两种权重方法，specificity和generality，边缘和节点越少，权重方法分配的specificity值越高，相反generality方法会给使用频率更高的边缘分配较高的权重值，Table1呈现了两种方法的权重值.
从表中可以看出，在specificity方法中Cook和Gleberman实例有最大的权重值，因为他们被使用的最少，因此boardMember关系属性最大。相反，Metaweb Technology实例比internet具有更多关联，但是在generality方法中internet实例有更高的权重。
Table3展示了与Table2相同条件下但是有考虑了边缘和节点权重的扩散激活输入的搜索结果，阈值自动设为0.01，Solicon Valley Colocation的节点权重比Metaweb Technology的更大，因此他排在第一，所以，Joseph Gleberman排第二尽管它距离初始节点更远，因为我们不考虑距离约束，从结果中看出，我们使用考虑权重的扩散激活得到了更有效的结果.

Table 1
Weight values of edges and nodes.

Edge/node	Weight value	
	Uniqueness	Generalness
includedCompany	0.367	0.494
companyFounder	0.215	0.839
boardMember	0.863	0.165
Internet	0.223	0.933
Science	0.664	0.375
Metaweb Technology	0.134	0.861
Silicon Valley Colocation	0.390	0.455
Applied Minds	0.223	0.681
Robert Cook	0.901	0.160
John Giannandrea	0.300	0.572
Danny Hills	0.570	0.322
Kevin Harvey	0.570	0.322
Joseph Gleberman	0.901	0.160

Table 2
Search results using edge weight based on uniqueness and the activation constraint.

Ranking	Results	Activation value
1	Silicon Valley Colocation	0.351
	Metaweb Technology	0.351
2	Joseph Gleberman	0.294
	Kevin Harvey	0.294

Table 3
Search results using edge and node weights based on uniqueness and the activation constraint.

Ranking	Results	Activation value
1	Silicon Valley Colocation	0.142
2	Joseph Gleberman	0.110
3	Metaweb Technology	0.049
4	Kevin Harvey	0.024

Result

Table4使用距离约束的排列前五结果，他们的激活值也被列出，在激活过程中，激活方法考虑了系统中每个节点，由于节点数较少，Silicon排第一，Metaweb排第二由于距离约束。
最后，Table5给出了使用了generality方法的距离约束扩散激活方法，搜索结果按照关联数目进行排序。
我们讨论了权重方法和扩散激活的影响，这个实验的目的是为了对比本方法相对于其他语义检索方法的优势和不足，我们选取了两种其他方法，SemRank和Rss。为了评价这个系统，我们参照rss的实验方法实施了一个调查，我们选取了五个测试查询分别用三种方法得到top10的查询结果。然后我们请了十个语义网方面的专家（语义网方面的硕士或者博士），他们中五个是我们实验室的，其他的是其他的成员，每个专家给每个搜索结果进行评分（0,1），0代表很不相关，0.3代表稍微相关，0.6代表相当相关，1代表高度相关，我们得到他们评分的平均数来评价每个方法的有效性 $sp(v_i) = \frac{\sum_{s=1}^{10}\sum_{j=1}^{k} score(v_i, s, j)}{10 \times k}$,
图5是评价结果，可以看出我们的generality方法超出了SemRank和Rss，可以说明用户搜索是更倾向于找到关联关系较多的，比较popular的信息。

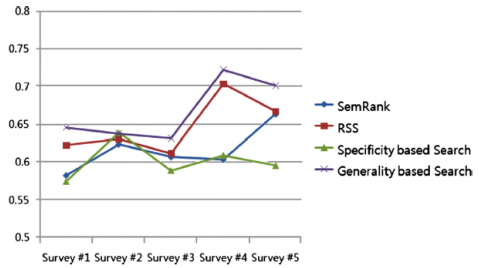

Fig. 5. Comparison of the scored precision.

Conclusion

本文中，我们提供了一种基于扩散激活的语义检索的方法，为了实现该目标我们提出了衡量自信息的方法，该方法考虑了属性和资源在语义网中的独特性，我们也提出了另一个基于generality的权重方法，来判断受欢迎概念。实验结果也显示这种方法是有效地，此外我们发现扩展激活算法的语义路径在语义网中更有价值。扩展激活方法可以得出与给定概念关联较强的概念集即便是概念建不存在知识base。

未来展望：我们计划限定度量标准来衡量语义关系，并将我们的方法应用于社会网络，此外我们将尝试开发考虑节点权重和相关关键词的初始激活值方法。最后，我们会建立一个语义检索网站来评价我们的方法。

谢谢指导！