Bilgi Çıkarımı
(Information Extraction-IE)

Doç.Dr.Banu Diri

Akış

- Bilgi çıkarımı nedir ?
- Mesaj anlama konferansları
- Uygulama alanları
- Yapılandırılmış, yarı yapılandırılmış dokümanlar
- Basit çıkarım şablonları
- NLP'nin bilgi çıkarımına katkısı
  - Öğelerine ayrılmış metinler
  - Özel anlamlı kelimeleri belirleme (Name Entity Recognition)
- Kaynak seçimi
- Dinamik web sayfalarından bilgi çıkarımı
  - Alışveriş robotları (froogle)
- IE performansının ölçümü
- Bilgi çıkarımında makine öğrenmesi
  - Şablonlar metodu için bir deneme

Bilgi Çıkarımı Nedir ?

- Yapılandırılmamış ya da yarı yapılandırılmış dokümanlardan önceden tanımlanmış şablonlara uygun bilgileri bulma
- Yapılandırılmamış ya da yarı yapılandırılmış dokümanların yapılandırılmış veri tabanlarına dönüştürülmesi

Mesaj Anlama Konferansları
Message Understanding Conference (MUC)

- Amerikan savunma bakanlığı 1990'lardan itibaren bilgi çıkarımı konusuna eğilmiştir.
- MUC her sene yapılan bilgi çıkarımı yarışmasıdır.
- Haber makalelerinden
  - Terör olayları
  - Şirketler dünyasındaki birleşmeler, yönetim değişiklikleri

  konularında bilgi çıkarımı

Uygulama Alanları

- İş ve işçi bulma
- Ürün bulma
- Seminer duyuruları
- Şirket bilgileri
- Üniversite başvuru bilgileri
- Kiralık / satılık daire, araba bilgileri
- *Ortak özellik ?*

birden fazla bilgi kaynağının araştırılması gereken durumlar

Yarı Yapılandırılmış Doküman
Örnek İş İlanı

Subject: US-TN-SOFTWARE PROGRAMMER
Date: 17 Nov 1996 17:37:29 GMT
Organization: Reference.Com Posting Service
Message-ID: <56nigp$mrs@bilbo.reference.com>

SOFTWARE PROGRAMMER

Position available for Software Programmer experienced in generating software for PC-Based Voice Mail systems. Experienced in C Programming. Must be familiar with communicating with and controlling voice cards; preferable Dialogic, however, experience with others such as Rhetorix and Natural Microsystems is okay. Prefer 5 years or more experience with PC Based Voice Mail, but will consider as little as 2 years. Need to find a Senior level person who can come on board and pick up code with very little training. Present Operating System is DOS. May go to OS-2 or UNIX in future.

Please reply to:
Kim Anderson
AdNET
(901) 458-2888 fax
kimander@memphisonline.com

Elde edilen iş özeti

computer_science_job
id: 56nigp$mrs@bilbo.reference.com
title: SOFTWARE PROGRAMMER
salary:
company:
recruiter:
state: TN
city:
country: US
language: C
platform: PC \ DOS \ OS-2 \ UNIX
application:
area: Voice Mail
req_years_experience: 2
desired_years_experience: 5
req_degree:
desired_degree:
post_date: 17 Nov 1996

Yapılandırılmamış Doküman Örnek Haber Metni

- 21 yaşındaki inşaat işçisi Kemal Yaprak, evine dönerken para meselesi yüzünden tartıştığı arkadaşı Hilmi Baker tarafından bıçaklanarak öldürüldü.

- Katil: Hilmi Baker
- Kurban: Kemal Yaprak
- Sebep: Para meselesi
- Suç aleti: Bıçak

Yapılandırılmış Doküman - Amazon Kitap Sayfası

```
….
</td></tr>
</table>
<b class="sans">The Age of Spiritual Machines : When Computers Exceed Human Intelligence</b><br>
<font face=verdana,arial,helvetica size=-1>
by <a href="/exec/obidos/search-handle-url/index=books&field-author=
        Kurzweil%2C%20Ray/002-6235079-4593641">
Ray Kurzweil</a><br>
</font>
<br>
<a href="http://images.amazon.com/images/P/0140282025.01.LZZZZZZZ.jpg">
<img src="http://images.amazon.com/images/P/0140282025.01.MZZZZZZZ.gif" width=90
  height=140 align=left border=0></a>
<font face=verdana,arial,helvetica size=-1>
<span class="small">
<span class="small">
<b>List Price:</b> <span class=listprice>$14.95</span><br>
<b>Our Price: <font color=#990000>$11.96</font></b><br>
<b>You Save:</b> <font color=#990000><b>$2.99 </b>
(20%)</font><br>
</span>
<p> <br>…
```


Elde edilen kitap bilgileri

Title: The Age of Spiritual Machines :
When Computers Exceed Human Intelligence
Author: Ray Kurzweil
List-Price: $14.95
Price: $11.96
:
:

Basit Çıkarım Şablonları

- **regular expression**
  - Amazon liste fiyatı:
    - $43.16
    - öncül şablon: "List Price: "
    - şablon: "\$\d\.\d{2}"
    - devam şablonu: ""

NLP'nin Bilgi Çıkarımına Katkısı

- Bilgiler dinamik web sayfalarından çıkarılacaksa basit regex şablonları yeterli olabilir.
- Bilgiler doğal, insanlar tarafından yazılmış metinlerden çıkarılacaksa NLP metotları yardımcı olabilir.
  - Part-of-speech (POS) tagging
    - Kelimelerin türünü (isim, fiil, sıfat vb.) belirleme
  - Sentaktik çözümleme
    - Kelime gruplarını, ağaçları belirleme, öğeleri bulma: NP, VP, PP
  - Anlamsal Kelime Sınıfları (WordNet'den)
    - KILL: kill, murder, assassinate, strangle, suffocate
  - Name Entity Recognition
- Örnek ***Öldürülen*** şablonu:

  Bart killed Rose.
  - Öncül şablon: [POS: V, synset: KILL]
  - Şablon: [Phrase: NP]

Öğelerine ayrılmış metinler

- "ye" fiilinin nesneleri yiyecek olarak sınıflandırılabilir.

geyiq.com/forum - Taksim borsada bira ile yarım döner **yerken**
geyiq.com/forum > geyiq alanı > kıl oluyorum, dumur oldum > Taksim borsada bira ile yarım döner **yerken**. Orjinalini görmek için **...**
www.geyiq.com/forum/archive/index.php/t-10219.html - 3k - Önbellek - Benzer sayfalar

Yerken Family Grave Search
... It's like we've always known each other!". - Pam from CA. Advertisement. Click Here. Search Page for Surname: **Yerken**. Name: First, Middle, **Yerken** Last. **...**
www.findagrave.com/surnames/y/**yerken**.html - 13k - Önbellek - Benzer sayfalar

Hürriyetim
... Kelebek, 25.05.2004. Ödülü, Bush kraker **yerken** söylemeyin, **...** Umarım kimse ona bu ödülü kazandığımı, o kraker **yerken** söylemez' diye yanıtladı. **...**
www.hurriyetim.com.tr/ haber/0,,sid~436@nvid~416896,00.asp - 42k - Önbellek - Benzer sayfalar

MİLLİYET İNTERNET - BUSINESS
... Zeytin **yerken** alzheimer oluyoruz haberimiz yok. Zeytini, zehirli tekstil boyası ile veya demir sülfat gübresi ile karartıp satıyorlar. **...**
www.milliyet.com.tr/2003/12/12/business/bus07.html - 30k - Önbellek - Benzer sayfalar

TurkiyeOnLine.com - Haber
... sağlık. Mantar **yerken** dikkat Havaların ısınmasıyla birlikte doğada ortaya çıkan mantarların bilinçsiz olarak tüketilmesinin, zehirlenmelere neden **...**
www.turkiyeonline.com/haber/saglik/ haber.php?story=2004_04_02_mantar - 20k - Önbellek - Benzer sayfalar

Özel anlamlı Kelimeleri Belirleme
Name Entity Recognition

- *NER sistemleri özel isimleri, tarih, yer, zaman ifade eden kelimeleri vs. belirlerler.*

- *Örnek:*
  - ***Jack Brown saw a cat in London.***
  - **[PER Jack Brown] saw a cat in [LOC London]** .
- *PER -- Person*
- *LOC -- Location*
- *ORG -- Organization*

- http://l2r.cs.uiuc.edu/~cogcomp/eoh/nedemo.html

Kural tabanlı NER sistemleri

- Yüksek performans
- Yüksek maliyet
- Kişi, şirket, yer isimleri listelerine ihtiyaç

If FullString = "New York" → LOCATION
If FullString = "California" → LOCATION
If Contains("Mr.") → PERSON
If Contains("Corp.") → COMPANY
If Contains("Inc.") → COMPANY
If Contains("Co.") → COMPANY
If Contains("Michael") → PERSON
If FullString = "Jordan" → LOCATION

- Şablon Örnek(ler)

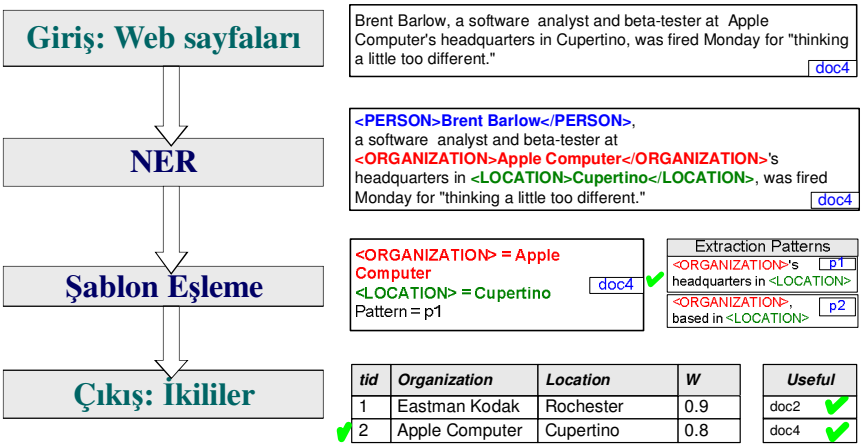
Examples of Keywords:
person titles (e.g., Mr., Jr., Ph.D.)
company designators (e.g., Corp., Inc., Co.)

- {TITLE}{PERSON}
  Ex: *"U.S. President George Bush", "Mr. Frank Leonard"*
- {PERSON}, the {TITLE} of {ORGANIZATION}
  Ex: *"Fred Martin, the CEO of XYZ Corp."*
- {PERSON} joined {COMPANY}
  Ex: *"Mary Smith joined Microsoft."*
- headquarters in {LOCATION}
  Ex: *"headquarters in London"*
- {LOCATION},{LOCATION}
  Ex: *"Salt Lake City, Utah"*

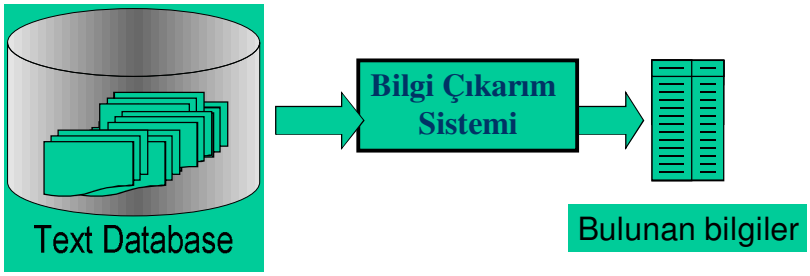
Tarih, telefon numarası, e-mail gibi kavramlar özel formatlarından tanımlanabilirler.

- Phone Number : (###) ###-####
- Date : ##/##/#####
- URL : www.xxxxxx.xxx/xxx/xxxx.html
- Email Address : xxxxx@xxx.xxxxx.xxx

NER ile Bilgi Çıkarımı



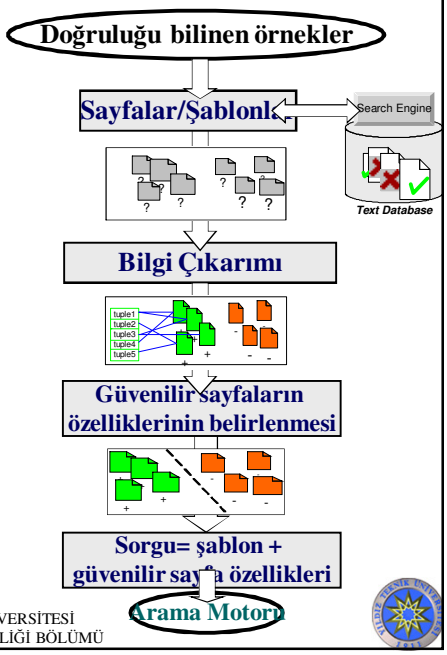
Kaynak Seçimi



- Problem: Bulunan bilgilerin güvenilirliği.
- Problem: Tüm web sayfalarını işleyemeyiz. Çok zaman alıcı.
- Çözüm: Sadece güvenilir Web sayfalarını işleyelim. Ama nasıl ?

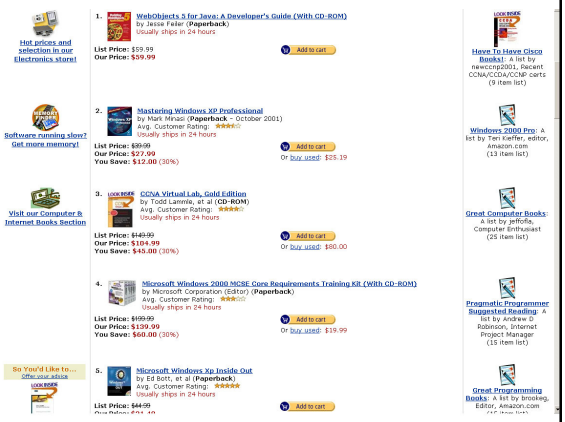
Kaynak Seçimi

- Doğruluğu bilinen örnekler arama motoruna gönderilir.
- Sonuçlardan şablonlar çıkarılır.
- Bu şablonlar arama motoruna gönderilir, uyan sayfalardan bilgiler çıkarılır.
- Çıkarılan bilgilerin doğruluğu bir veritabanından kontrol edilir.
- Doğru ve yanlış bilgi çıkarılan web sayfaları işaretlenir.
- Bu sayfaların özellikleri çıkarılır.
- Bundan sonra şablonlarla birlikte doğru sayfaların özellikleri de aratılır.
- Bu sayede sadece güvenilir sayfalarda arama yapılmış olunur.
- ÖZETLE: Bulunan şablonlara ek olarak, güvenilir sayfaların özellikleri de bulunarak sorguya eklenir.
- SAYFA ÖZELLİKLERİ: İçinde geçen kelimeler, url'sinde geçen kelimeler (ör: *edu*)



Dinamik Web Saylarından Bilgi Çıkarım Metotları

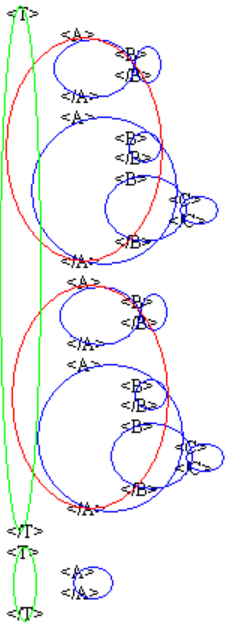
- Birçok web sayfası veritabanlarından dinamik olarak oluşturuluyor.
- Dinamik web sayfalarında html tag'leri tekrar eder.
- Tekrar eden kalıplar arasında aynı tür bilgiler yer alır.



Tablomuzun Satırlarını Belirlemek

Satırlar başlayıp biten HTML tag'lerinden oluşur.
Hangi tag'le satırın başlayıp bittiğini bulmak önemli.

Kural 1: Her satırdaki HTML tag sayısı birbirine yakındır/eşittir.
Kural 2: En fazla tag içeren tekrarlı çevrim satırı gösterir.



- Yanda olası tüm satırlar gözükmektedir.
- Her satırda yakın sayıda tag olması şartından dolayı T'lerin satırları oluşturmadığı görülür.
- En fazla sayıda tag içeren satır seçileceğinden kırmızı ile gösterilen kısımlar satırlar olarak belirlenecektir.

Alışveriş Robotları

- Tekrarlı HTML tag'leri kullanılarak bilgi çıkarılan sistemlere örnek olarak çeşitli web sitelerinde satılan ürünlerin bilgilerini tek bir sayfada toplayan sistemler verilebilir.
- Örnek Siteler:
  - MySimon
  - Cnet
  - BookFinder
  - Froogle

Alışveriş/Haber Toplama Robotlarının Çalışma Adımları

- 1- Her satıcı/haberci site bilgi çıkarım mekanizmasını kurar.
- 2- Kullanıcıdan sorgusunu alır (tür, fiyat vs.).
- 3- Her site için:
  - Kullanıcı sorgusu siteye gönderilir.
  - Sonuç sayfaları alınır.
  - Sonuç sayfası, o sayfanın bilgi çıkarım mekanizmasıyla işlenir. Sonuçlar kendi veritabanına kaydedilir.
- 4- Sonuçlar (fiyatlara/tarihlere göre) sıralanır.
- 5- Sonuçlar HTML formatına çevrilir. Kullanıcıya döndürülür.

IE performansının ölçümü

- Performans, sistemin eğitimi sırasında kullanılmamış olan elle işaretlenmiş test verisi üzerinde ölçülür.
  - Dokümanlarda yer alan doğru cevap sayısı: N
  - Sistem tarafından çıkarılan toplam cevap sayısı: E
  - Sistem tarafından çıkarılan toplam doğru cevap sayısı: C
- Ölçütler
  - Recall = C/N
  - Precision = C/E
  - F-Measure = 2*Recall*Precision/(Recall+Precision)

Şablonların bulunması

- Keşfetmek istediğimiz ikililerin aralarındaki ilişki türü belirlenir. Ör: "Tüm X'ler Y'dir".
- Bilinen X,Y ikilileri Google'da aratılır.
- X ve Y arasındaki şablonlar ve frekansları belirlenir.
- En yüksek frekansa sahip şablonlar bu ilişki türünün şablonları olurlar.

Bulunan şablonlardan örnekler tüm X'ler Y'dir için

- ve diğer
- ler ve diğer
- ve benzeri
- veya diğer
- türü olan
- ları ve diğer
- lar ve diğer
- ve her türlü
- lerden biri olan
- leri ve diğer
- larından biri olan
- lerinden biri olan
- lardan biri olan
- adı olan
- ve her tür

Bulunan şablonlardan örnekler
X'in yeri Y'dir için

- y deki x
- y de bulunan x
- y de x
- x y de
- x y ili sınırları içerisindedir
- y ili sınırlarında kalan x
- y ili sınırları içinde bulunan x
- y ilçesi sınırları içinde bulunan x
- x y nin sınırları içerisindedir
- x/y
- x / y
- x-y
- x y ye zz km

x, y ye zz km
x (y ye zz km
x, y
x - y
x bulunduğu yer:y
y-x
x(y
x(y)

Şablonlara uygun ikililerin bulunması

- Google'da bulunan şablonlar aratılır.
- Sonuç sayfalarındaki şablonların sağ ve sollarındaki kelimeler alınır ve bir dosyaya kaydedilir.

Şablonlara uygun ikililerden örnekler

- Tüm X'ler Y'dir
- kontrolör personel
- teçhizat malzeme
- kemer teçhizat
- protein gıda
- Azerbaycan bölge
- Ceyda yardımcı
- komünizm ideoloji
- delta Gediz
- kurum Kocaelispor
- fotoğrafçı Robert
- tür flamingo
- ünite aksesuar
- bedel masraf
- din azınlık
- çelik yapı
- yem araç
- kız sıfat
- yapı sorun
- ölçü şart

İkililerin elle sınıflandırılması

- Bulunan ikililerden hangilerinin "Tüm X'ler Y'dir" ilişkisine sahip olup olmadığı elle işaretlenir.

Kaynaklar

- Rada Mihalcea, “NLP lecture slides”
- www.ccs.neu.edu/home/futrelle/ bionlp/psb2001/Hawaii-Tutorial-Tsujii.ppt
- www.cs.utexas.edu/users/mooney/ ir-course/slides/InformationExtraction.ppt
- www.cs.columbia.edu/~eugene/talks/icde2003.ppt
- www.isi.edu/natural-language/ teaching/cs544/cs544-9-apr04.ppt
- www.cs.sfu.ca/~zshi1/personal/ projects/Presentation_thesis.ppt