Türkçe Metin Arama Sistemleri ve Veritabanlarının Karşılaştırılması

Özgür Yılmazel

Anadolu Üniversitesi Bilgisayar Mühendisliği

İstanbul, Nisan 2009



Sunu Akışı

- ☐ İnternet Uygulamaları ve LAMP
- Metin Arama
- Veritabanı ve Metin Arama Sistemleri
- Değerlendirme
- Uygulama Örnekleri



İnternet Uygulamaları

- e-ticaret uygulamaları, bilgi portalları
 - **İ**çindekiler:

Veritabanı, web sunucusu, sunucu üzerinde çalışan yazılımlar, ve (html+javascript) kullanıcı arayüzleri

LAMP – Açık kaynak kodlu çözüm

- Linux, Apache, MySQL, (PHP|Perl|Python)
- Tüm dünyada çok hızlı bir şekilde benimsendi
- Düşük maliyet ile çok becerikli uygulamalar geliştirildi
- Düzenli veri yapılarının saklanması ve tekrar ulaşılması için güçlü çözüm

İnternet Uygulamaları – Düzensiz Veriler

- Düzenli Veri Fiyat, Tarih, Miktar SQL ile aranılabilir
- ☐ Düzensiz Veri Metin, resim ve video
 - Metinsel veriler:

□Email mesajları	□İş ilanları
□Ürün açıklamaları	☐Haber metinleri
□Raporlar	

- Kurumların sahip oldukları düzensiz veri miktarı büyük bir hızla artmakta "Şirketlerin sahip oldukları verinin %80i düzensiz veridir." Sun Data Warehouse Reference Architecture, Mart 2008
- "2009 yılının sonunda 4 exabyte bilgi oluşturulacak bunun 3 exabyteından daha fazlasıda düzensiz bilgi olacaktır. Ve bu miktar son 5000 yılda oluşan bilgiden fazladır" Karl Fisch
- Bilgi ihtiyacının karşılanabilmesi için düzenli ve düzensiz verilerin tümünde efektif bir ulaşım mekanizmasına ihtiyaç var.



Veritabanları ve Düzensiz Veriler

- Geleneksel Veritabanları ve arama dili SQL düzensiz verilerin saklanması ve aranması için uygun değil.
- ☐ Metinsel veriler için arama sistemleri (Information Retrieval Systems)
 - Sıralama (ranking)
 - Sorgulama dili
 - Eşleştirme
- Bir çok öncü firma kendi arama sistemlerini geliştirmek için yatırımlar yaptılar:
 - Amazon A9
 - monster.com Temmuz 2008'de Trovix arama şirketini \$72.5Milyon satın aldı
 - e-Harmony Aşık olma olasılığının matematiksel modelini oluşturdu
 - eBay 1995'de kuruldu, 1998'de kendi arama motorunu geliştirdi



Metin Arama Sistemleri

Veritabanı üreticisi firmalar yazılımlarına tam-metin indeksleme ve arama özellikleri eklemeye başladılar. (Değişen başarı standartlarında)

IBM DB2 – Net Search Extender	1983	1999 (5)
Microsoft SQL Server	1989	1998 (7.0)

Oracle – Ultra Search Text Search 1971 1999 (8i)

MySQL 1995 2000 (3.23)

PostgreSQL 1989 2001 (7.0.3)

- ☐ Veri depolama amacı güdülmeden, sadece bilgiye ulaşımın kolaylaştırılması ve arama başarımının arttırılmasına yönelik sistemler
 - Apache Lucene
 - Tamamen Java'da yazılmış, açık kaynak kodlu, yüksek performanslı ve ölçeklenebilir bir metin arama kütüphanesi.
 - Bazı kullanıcıları: CNET Reviews, Expert Exchange, Wikipedia, IBM OmniFind



Türkçe Metin Arama ve Veritabanı Karşılaştırma Motivasyonu

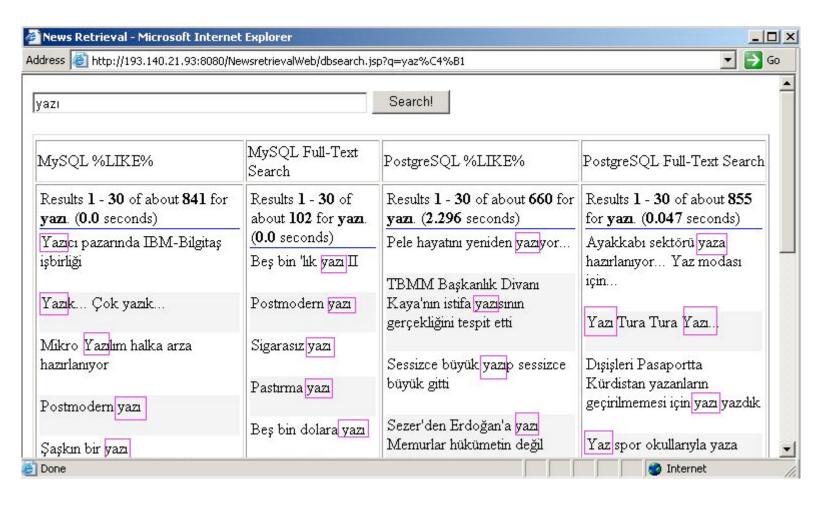
- Türk kurum ve kuruluşları ellerindeki metin verilerinden tam anlamıyla faydalanamamakta ve bu verilerden yarar sağlayamamaktadır.
 - e-ticaret uygulamalarında aranılan ürünlerin bulunamaması
 - Güncel bir gazetenin arşivinde arama yaptığınızda doğru habere ulaşamamanız
 - Bir doktorun veya hukukçunun doğru güvenilir bilgiye kolay ulaşamaması
- ☐ Türkçe Metin Arama konusunda yapılan çalışma sayısı oldukça az
- Yapılmış olan çalışmaların ticari uygulamaya dönüşmüş örnekleri yok veya yaygın olarak bilinmiyor.
- Veritabanları tam metin indeksleme ve arama özellikleri geliştirseler de, bu özellikler ingilizceyi hedef alıyor
- Türk firmaların çoğunluğu LAMP içinde sıkışmış durumdalar
 - daha iyi bir yol var!

En sık kullanılan yöntem

"tüm doküman bir string olsun".indexOf("sorgu") şeklinde yapılan aramalar bugün elimizde olan veri kalabalığında çok ilkel kalıyor.



Veritabanları Arama Sonuçları





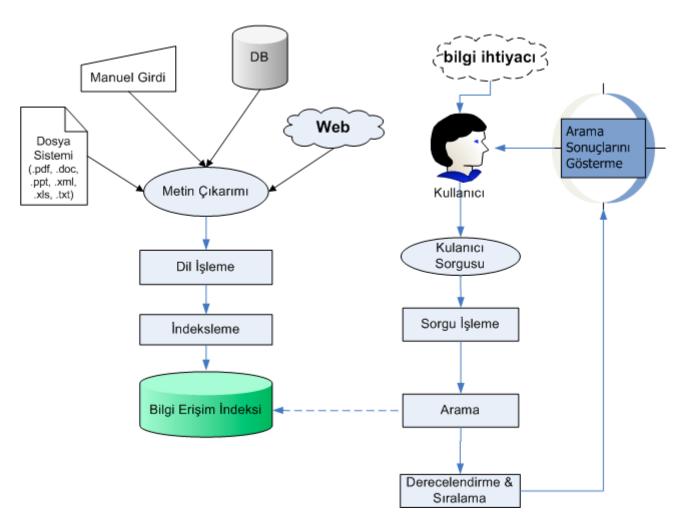
Amaç

Açık kaynak kodlu yazılımları kullanarak Türkçe için uygun, ticari olarak kullanılabilir bir arama sistemi gerçekleştirmek, bu sistemin mevcut veritabanına dayalı sistemler ile karşılaştırmasını yapmak.

Ticari olarak kullanılabilirlik:

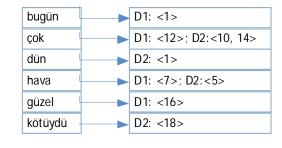
- ☐ Hızlı indeksleme ve sorguya cevap verme süresi
- Ölçeklenebilir tepki zamanının indeks boyutuna göre değişimi (binlerce, milyonlarca doküman üzerinde arama yapabilme)
- Anlamlı sonuçlar İndexde var olan sonuçların eksiksiz ve doğru sıralamada kullanıcıya sunulması.







- Indeksleme
 - Hızlı arama için özel veri yapıları
 - D1 Bugün hava çok güzel.
 - D2 Dün hava çok çok kötüydü.



Dictionary

Posting List



- Arama: ters indeks dosya yapıları ile hızlı arama
 - Arama Özellikleri:
 - Normal kelime arama
 - Tamlama Aramaları "kuş gribi"
 - Bulanık Aramalar biodizel~ → biodizel

biyodizel

bio-dizel

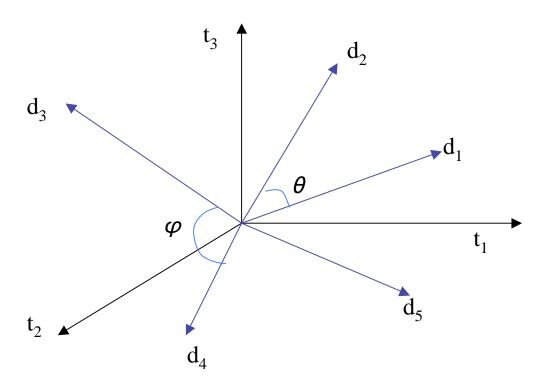
- Yakınlık Arama "anayasa demokrasi"~25
- Eş anlamlı arama doktor → doktor

hekim

- Sonuçların anlamsal olarak sıralanması
 - Veritabanlarında yapılan metin sorguları küme döndürür
 - Anlamsal yakınlığa göre sıralı sonuçlar

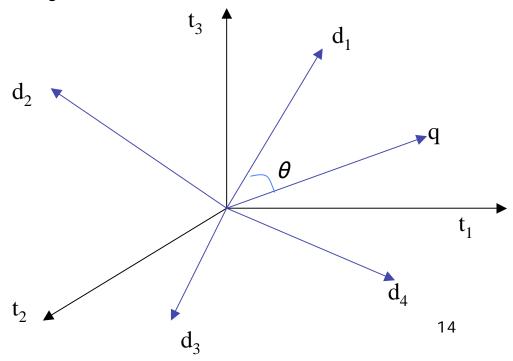
Metin Arama – Vektör Uzayı Modeli

- Dokümanlar ve sorgular t-boyutlu vektörler
- Vektör elemanları farklı ağırlıklandırma yöntemleri ile belirlenir
- Dokümanların sorgulara olan yakınlıkları bu ağırlıklar ile hesaplanır



Metin Arama – Kosinüs Benzerliği

- Benzerlik(dj,q)= $\cos \theta$
- Sonuçlar benzerliğe göre sıralanıp kullanıcıya gösterilir
- Doküman uzayının boyutu tüm dokümanlarda geçen ayrı kelime sayısına eşittir (400Bin dokümanlık İngilizce bir doküman seti için ~30Bin boyut)





Metin Arama – Dil İşleme

- Benzerlik ölçüsünün en önemli dayanağı anlamlı metin uzaylarıdır, verilen doküman kümesinden metin uzayının boyutlarının belirlenmesi sırasında değişik dil işleme metotları kullanılır.
 - Uzayın boyutlarının azaltılması indeksin boyunu küçültür
 - Arama hesaplamalarında hız kazanmamızı sağlar

run, ran, running

- Stopword Eliminasyonu sık kullanılan ve anlam içermeyen kelimelerin indekse alınmaması; ve,veya, eğer
- Gövdeleme Birbirinin türevi olan kelimelerin tek bir boyutta toplanması student, students



Türkçe Arama

- Türkçe
 - Sondan eklemeli bir dil
 - Sesli uyumu var
 - Cümle bileşen sırası serbest

Türkçedeki kelimelerin büyük çoğunluğu oldukça küçük temel bir sözlükteki köklerin sonlarına yapım ve çekim ekleri eklenerek türetilmiştir.

- ~400Bin dokümandan oluşan Türkçe bir doküman setinde
- ~1.3 Milyon ayrı kelime, boyut vardır.

Türkçe Metin Arama için Dil İşleme

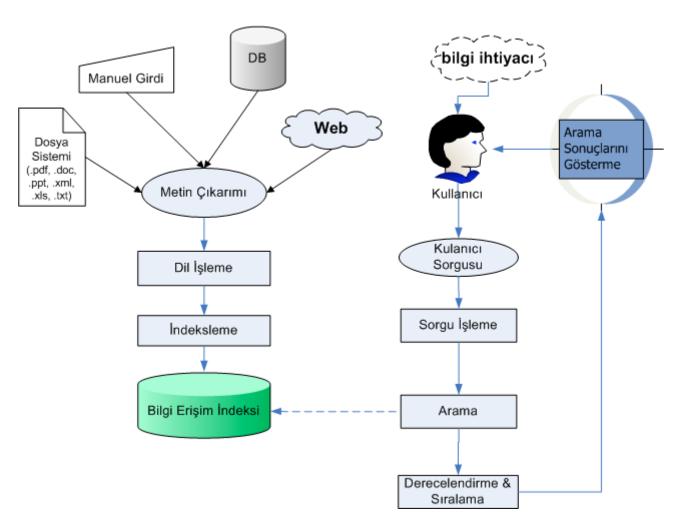
- ☐ Gövdeleme (4 farklı sistem denendi)
 - Gövdelemesiz + stopword: out-of-the-box settings
 - Snowball Gövdeleme:

 (Kelimeleri kök hallerine indirger)

 org.apache.lucene.analysis.snowball.SnowballAnalyzer
 - Ilk 5 Karakter Gövdeleme: (F.Can et.al JASIST 2008) Tüm kelimelerde sadece ilk 5 karakteri kullanır
 - Zemberek Tabanlı Gövdeleme:

http://code.google.com/p/zemberek





- MySQL
- PostgreSQL
- Lucene

MySQL

MySQL tam-metin indekslemeyi ve tam-metin aramayı destekler; fakat, MySQL'in tam-metin arama fonksiyonlarının herhangi bir dil için, dile özel dilsel desteği yoktur. Sadece İngilizce için stopword eliminasyonu mevcut.

- PostgreSQL
- Lucene

- MySQL
- PostgreSQL

PostgreSQL, OpenFTS(Açık Kaynaklı Tam Metin Arama Motoru) kullanır. OpenFTS verilerin online indekslenmesini ve veritabanı aramaları için ilgi sıralaması sağlayan gelişmiş bir arama motorudur. PostgreSQL dilsel destek açısından birçok dil için önceden tanımlanmış dil işleme metodları içerir, Türkçe de bu dillerden bir tanesidir.

Lucene

- MySQL
- PostgreSQL
- Lucene

Apache Lucene, tamamıyla Java dilinde yazılmış, yüksek performanslı ve ölçeklenebilir bir Bilgi Erişim(IR) kütüphanesidir. Tam-metin araması gerektiren, özellikle farklı platformlarda çalışacak, tüm uygulamalarda kullanılabilecek bir teknolojidir.

http://lucene.apache.org



Sistemlerin Karşılaştırlması

- Bilgi Erişim Sistemlerini değerlendirmek için öncelikle bir test seti gereklidir.
 - Test Seti:
 - Dokümanlar
 - Bilgi İhtiyaçları
 - Cevap Anahtarı (relevance judgements)
- ☐ Tekrar kullanılabilir test setleri oluşturmak çok zordur.
- Türkçe için tek test seti Bilkent IR Research Group'un oluşturduğu, TREC-benzeri Türkçe test seti olan "milliyet kolleksiyonu"



Milliyet Test Seti

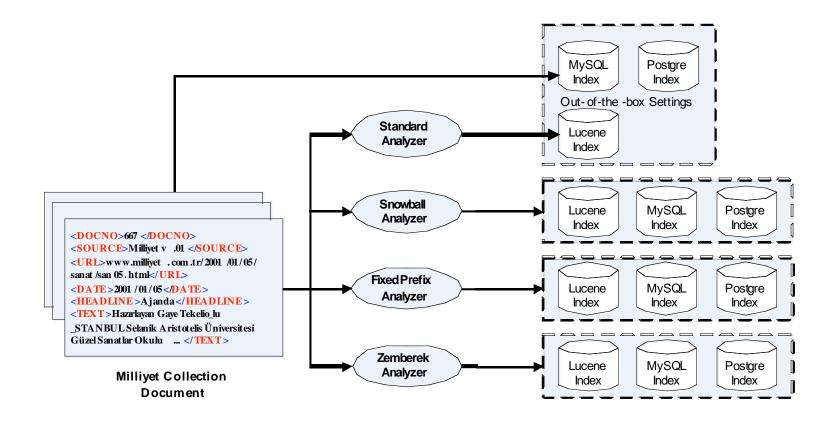
408,305 Doküman

Cevap Anahtarı

72 konu, sorgu

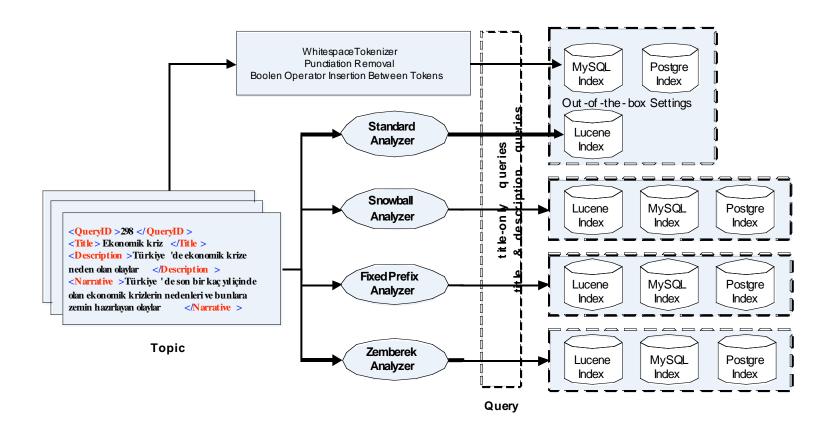


Deney Sistemi - İndeksleme





Deney Sistemi - Arama





Değerlendirme Ölçütleri

$$Precision = \frac{\#(relevantDocumentsRe\ trieved)}{\#(retrievedDocuments)}$$
10/20

$$Re \, call = \frac{\#(relevantDocumentsRe \, trieved)}{\#(relevantDocuments)}$$
10/30

Bilgi erişim sistemlerinde kullanılan standart ölçütlerdir.

Örnek: Toplam 100 doküman. "kuş gribi" sorgusu ile ilgili 30 doküman var. Sistem 20 doküman geri döndürdü, bunlardan 10 tanesi "kuş gribi" hakkında.



binary preference (bpref)

Yeni bir ölçüt olan bpref eksik cevap anahtarlı bilgi erişim sistemlerinin karşılaştırılmasında kullanılmak üzere tasarlanmış bir ölçüttür.

Chris Buckley ve Ellen Voorhees trec_eval yazılımı

bpref sadece cevap anahtarında bulunan ilgili dokümanların ilgisiz dokümanlardan daha yukarı sıralarda olması ilişkisi üzerinde hesaplanır. Sistemlerin tüm set üzerindeki performanslarının güçlü bir ölçütü olmasada, iki sistemin karşılaştırılabilmesi için iyi bir ölçüttür.



Deney Sonuçları (bpref)

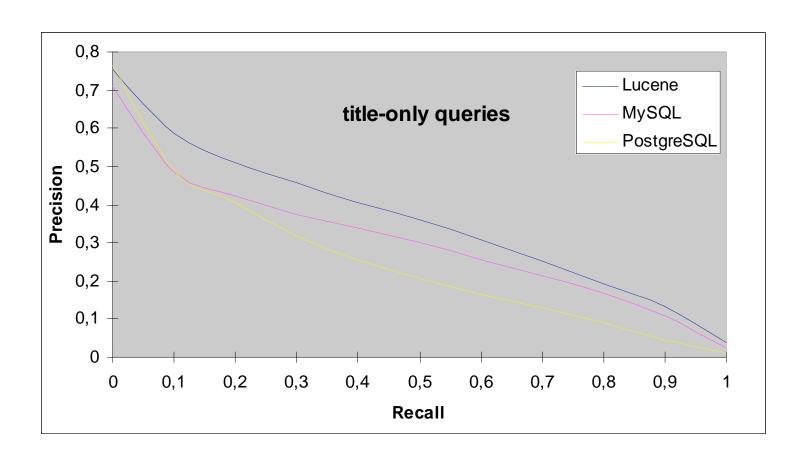
Sadece başlık Aramaları	Standart	Snowball	ilk5	Zemberek
Lucene	0.3505	0.4313	0.4413	0.4506*
MySQL	0.2899	0.3388	0.3637	0.3644
PostgreSQL	0.4249	0.4187	0.4308	0.4389

Başlık ve Açıklama Aramaları	Standart	Snowball	ilk5	Zemberek
Lucene	0.3947	0.4867	0.5067	0.5124*
MySQL	0.3019	0.3808	0.4139	0.4140
PostgreSQL	0.2672	0.2932	0.2808	0.2660

^{*} Bu sonuçlar milliyet test seti için literatürde raporlanmış en yüksek sonuçlardır.

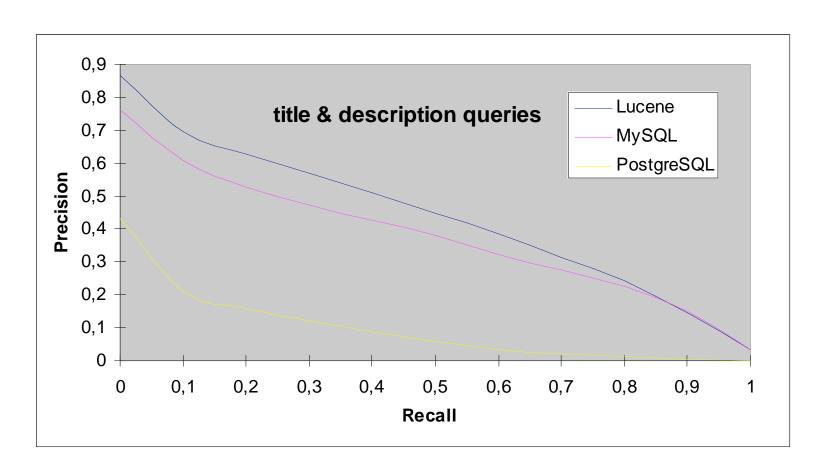


Precision – Recall Grafiği



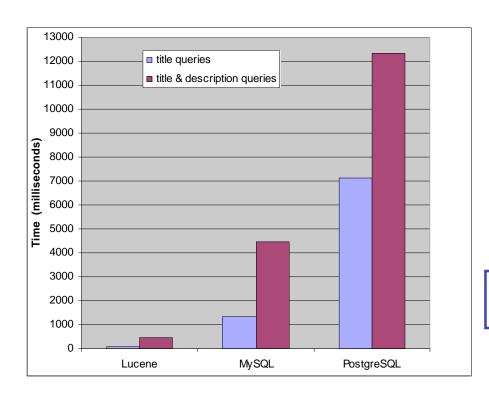


Precision – Recall Grafiği





Ortalama Arama Süreleri



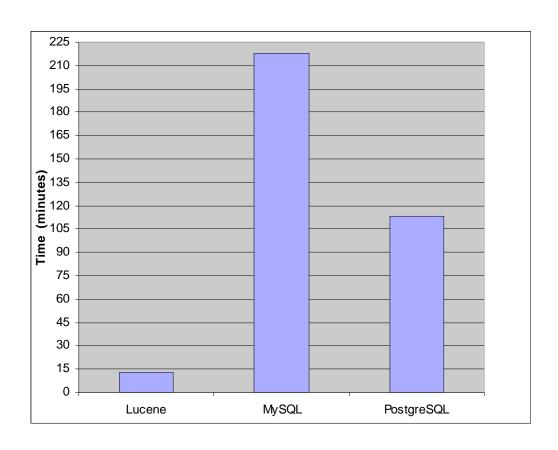
Gövdeleme	Ayrı kelime sayısı
Standart	1,346,183
Snowball	801,548
Zemberek	597,658
ilk5	262,347

Başlık sorguları: ortalama 3 kelime

Başlık ve Açıklama: ortalama 16 kelime



Ortalama İndeksleme Süreleri





Değerlendirme

- Türkçe değerlendirme kümesi için en etkili ve verimli sonuçların Türk diline özel geliştirilmiş bilgi erişim kütüphanesinin kullanılmasıyla elde edildiği gösterilmiştir.
- Turkish Analyzer herhangi bir dile özel gelişim içermeyen ilişkisel veritabanlarına göre performansı %55 arttırmıştır.
- Lucene'in indeksleme sırasındaki işlem zamanı ilişkisel veritabanlarına göre 10 kat daha iyidir. Arama sırasındaki işlem zamanı ise 50 kat daha iyidir.
- İlişkisel veritabanları kolay kullanılabilir tam metin arama özelliği sağlasada, dil işleme yapılmadığından performansları oldukça düşüktür.
- Aranan kelime sayısı arttıkça ilişkisel veritabanlarındaki arama hızı düşer. Metin Arama Sistemlerinde ise kelime sayısındaki artış hıza yansımaz.



Uygulama Senaryoları

- Sektöründe lider site:
 - ☐ Günde 500.000 sorgu
 - 2 Milyon Sorgu üzerinde yapılan incelemede
 - Mevcut sistemden ortalama 20 kat daha hızlı

	Ortalama	Maksimum
Mevcut Veritabanı	0.188 sn	53.1262 sn
BES	0.0095 sn	6.36 sn

- Mevcut Veritabanı 8 Core üzerinde
- ☐ Bilgi Erişim Sistemi tek core üzerinde



HaberTakip

http://www.habertakip.com

17 Haber kaynağından 1,233,142 haber

Haber Kümele (Clustering)



Günlük Haber Ara:

Gelişmiş Arama

gösteri, dönemin başbakanı İsmet İnönü'ye, o zamanki ABD Başkanı Johnson tarafından yazılan "saygısız yapıyor. 6-7 Eylül'ü yaşadığımda

Kaynaklar

Arama Sonuçları: 'ABD Başkanı' başlığı için yaklaşık 35 sonuçtan true 1-10 arasi sonuçlar 0.02 saniye

- Sabah (6)
- Haber7 (4)
- Hurriyet (4)
- Vatan (4)
- Haber3 (3)
- HaberTurk (3)
- InternetHaber (3)
- Milliyet (3)
- Samanyolu (2)
- EnSonHaber (1)
- Radikal (1)
- Star (1)

<u>®iviiiiyet</u>

Millivet

100 %

İran ve Ortadoğu 06 Ocak 2009, Salı

Vitrinlerde Filistin bayrakları 06 Ocak 2009, Salı

çocukluk çağının sonlarındaydım. Türkiye'de ilk kez...

100 % O

ABD'nin yeni Başkanı Barack Obama, "Change (Değiş)" sloganıyla kazandı. Henüz resmen başkanlık makamını devralmamış olsa da, Obama, bakanlarını seçti ve kilit bürokratlarını atadı. ABD'de...

Milliye

İsrail'i kim durduracak? 06 Ocak 2009, Salı

100 %

" diye açıklanamaz. Yeni ABD Başkanı da "susarak" İsrail'e destek veriyor. Washington'ın mevcut veya) veya bölge ülkelerinin İsrail'i durdurmaları mümkün değil. ABD, İsrail'in arkasında olduğu...

Katagoriler

- Baska (14)
- Dunya (11)
- Siyaset (4)
- <u>Ekonomi</u> (1)
- Gundem (1)
- Politika (1)
- Spor (1)
- Turkiye (1)
- Yazarlar (1)

Yıllar

2009 (35)

haber@com Polise en düşük 200 YTL zam geliyor 06 Ocak 2009, Salı

100 %

, polise maktu olarak fazla mesai ücreti de ödendiğini ifade etmişti. Ancak polis yine de ABD ve Avrupa da zam Türk polisi, Avrupalı ve ABD'li meslektaşlarına oranla daha az maaş alıyor. Polislerin emniyet müdürü,...

PORFR

100 %

Bush'tan Blair'e İstiklal Madalyası 06 Ocak 2009, Salı

İngiltere'nin eski başbakanı Tony Blair'in, ABD Başkanı George W. Bush'tan "Başkanlık İstiklal . Biri ABD Başkanı, diğeri ise ABD Kongresi tarafından verilen her iki madalya da ABD'de...

VATAN

Gazze için başka planlar! 06 Ocak 2009, Salı

ABD'de seçimden sonra Başkan'ın yetkilerinin sınırlı olmasına, böylece bir karara tek başına imza olunca yürek dayanmıyor. Aynen Irak'ta ABD saldırılarında ölen çocuklar gibi... Aynen PKK terörüyle...

100 %

100 %

CIA'de sürpriz 06 Ocak 2009, Salı

ABD'de, 20 Ocak'taki yemin töreninden sonra başkanlık görevini üstlenmesi öngörülen Barack Obama'nın, Merkezi Haberalma Örgütü CIA'in başına, ABD'nin eski başkanlarından Bill Clinton döneminin eski...

PORE

100 %

Nihayet konuştu! 06 Ocak 2009, Salı

ABD'de bu ay yemin töreninden sonra başkanlık görevini devralacak Barack Obama, Gazze'deki durumla koymadığı gerekçesiyle gelen eleştirilere karşılık, "ABD'nin bir tane başkanı var" dedi ve ABD...



100 %

Bahçeli bu hatayı nasıl yaptı? 06 Ocak 2009, Salı

çıkıyor. Bilinen ilk dinleme skandalı Watergate, ABD Başkanı Richard Nixon'ın istifasıyla sonuçlanmıştı ve istifa eden tek başkan olarak tarihe geçmişti. Bizim gündemimize, dönemin Başbakan Yardımcısı...

Google Reklamları

Antalya'da Havuzlu Daire 120 Ay Vadeli % 0,95 Sabit Faizli Kredi ile Denize Yakın Lüks Evler! www.Ekol.biz/Antalya_l

Canlı Borsa Haber

Hesap açana Foreks FX2000 veriyoruz İnternetten Hisse

Onbinde2 www.sankomenkul.com

Medya Takip

Şirketinizin ve Rakiplerinizin Tüm Haberlerine Hergün Online Erişin! www.prnet.com.tr

Kobi Dergi

Okuyarak değil, izleyerek öğrenin. TEB Kobi TV ile işinizi büyütün TEBkobitv.com/KOBI

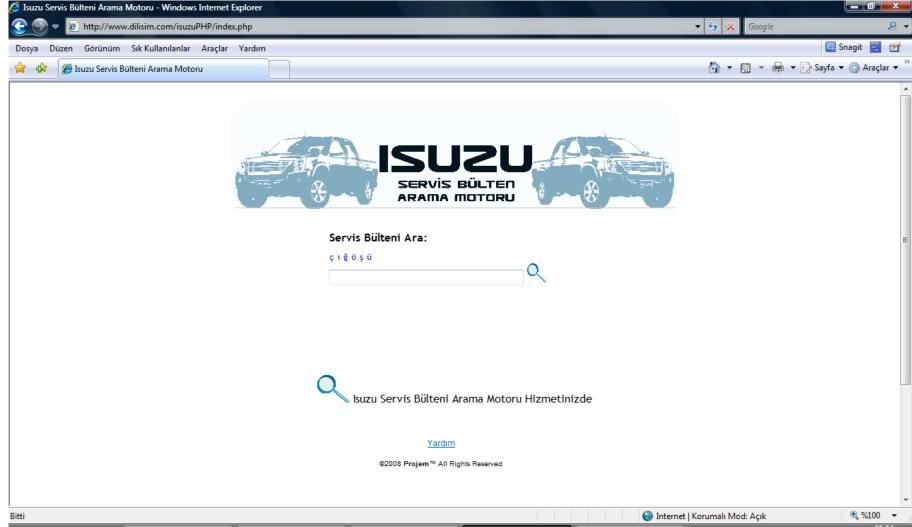
A V



Nuriş Oto Servis Bülten Arama Motoru

http://www.dilisim.com/isuzuPHP







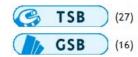
Servis Bülteni Ara:

Q

Ilgili Araç

D-MAX (2)
08.09.2006 ila
15.05.2007 arası
üretilmiş (1)
CITIBUS (1)
D MAX (TFS 77) (1)
DİSK FRENLİ
TURKUAZ VE CITIBUS
ARAÇLAR (1)

Bülten Çeşidi



Yıllar

2004 (7) 2005 (3) 2006 (3)

2007 (4) 2008 (26) Arama Sonuçları : fren balata için yaklaşık 43 sonuçtan 1 - 10 arasi sonuçlar (0.09) saniye



HAVALI FREN BALATA ve KAMPANA AŞINMASI 02 Haziran 2006, Cuma

KONU: HAVALI FREN BALATA ve KAMPANA AŞINMASI 12 Haziran 2006 tarihinden itibaren Turkuaz, MD serisi ve NQR 3D araçların fren balata - kampanaları ile ilgili claimler, diğer araçlarda



GSB04009 yenilik ek 07 Kasım 2008, Cuma

Otomatik fren balata ayarı ğ ğ MZZ 6U Şanzıman ğ MYY 5T Şanzıman ğ 7,5 ton toplam yüklü ağırlık ğ ğ Yenilikler NPR 66 NQR 70 ABS (Anti Blokaj Fren Sistemi) ğ ğ HBB Fren sistemi ğ ğ LSPV ğ ğ



TSB 0706 citibus EK1 07 Kasım 2008, Cuma

fren balataları 15mm kadar büyütülmüştür. Böylelikle daha fazla frenleme kuvveti sağlanmıştır. Ön geçilmiştir Ön aks rulmanları değiştirilmiştir Ön aks fren balataları için aşınma ikazı konulmuştur Direksiyon CITIBUS ARACI GENEL TANITIM NOTLARI Araç Tipi Yurtici satış şanzıman fren sistemi süspansiyon



GSB07032 EK-1 07 Kasım 2008, Cuma

) KOMPLE DEĞİŞİMİ (Fren balatalarının sökülmesi, temizlenmesi, kontrolü ve hava alınması dahil) FREN HİDROLİK BOOSTER DEĞİŞİMİ FREN PABUÇLARI VE / VEYA BALATALARI DEĞİŞİMİ (Tekerlek, kampana, porya ve fren ARKA AKS KOVANI DEĞİŞİMİ ÖN AKS MUHAFAZASI (KOVANI) DEĞİŞİMİ (4 x 4) DİSK FREN MERKEZİ (KALİPER



FREN AYAK VALFÍ HORTUM MONTAJ DEĞİŞİKLİĞİ. 02 Haziran 2005, Perşembe

. Bu değişikliğin amacı arka fren balatalarının ömrünü uzatmaktır. Tarafınıza gönderilen uygulama talimatına KONU: FREN AYAK VALFİ HORTUM MONTAJ DEĞİŞİKLİĞİ. İLGİLİ ARAÇ:TURKUAZ (1023-1979 şase numaraları arasında kalan araçlar.) Sayın yetkili servisimiz, Havalı fren sistemine sahip Turkuaz araçlarının fren



GSB04004 yenilik ek 07 Kasım 2008, Cuma

ğ ğ 9.0 Otomatik fren balata ayarı ğ ğ 10.0 Hız sınırlayıcı ğ 11.0 200 litrelik yakıt deposu ğ Tam havalı, çift devreli fren sistemi ğ ğ 4.0 ABS (Anti Blokaj Fren Sistemi) ğ ğ 5.0 ASR (Anti patinaj kontrolü) ğ ğ 6.0 LSV ğ 7.0 Arka tekerleklere bağlı el freni ğ ğ 8.0 Hava takviyeli imdat freni



GSB04004 yenilik X 07 Kasım 2008, Cuma

Otomatik fren balata ayarı ğ ğ ğ ğ ğ ğ ğ ğ ğ ğ Hız sınırlayıcı ğ 200 litrelik yakıt deposu ğ MZZ 6U tasarımı ğ Yeni kumanda konsolu tasarımı ğ Tam havalı, çift devreli fren sistemi ğ ğ ğ ğ ğ ABS (Anti Blokaj Fren Sistemi) ğ ğ ğ ğ ğ ğ ASR (Anti patinaj kontrolü) ğ ğ HBB Fren sistemi ğ ğ LSPV ğ ğ ğ ğ ğ



FREN KAMPANASI İÇ ÇAPI ÖLÇÜMÜ VE KONTROLÜ 02 Şubat 2006, Perşembe

KONU: FREN KAMPANASI İÇ ÇAPI ÖLÇÜMÜ VE KONTROLÜ İLGİLİ ARAÇ: TURKUAZ, NQR-3D, ROYAL, URBAN, CLASSIC Fren kampanalarının kontrolü ve tornalanması için aşağıdaki adımlar takip edilmelidir. Fren kampanısında aşırı aşınma ve çizilme olup olmadığını göz ile kontrol edin. Eğer fren kampanası aşınmış



Ön sol fren hortumları 21 Mayıs 2008, Carsamba

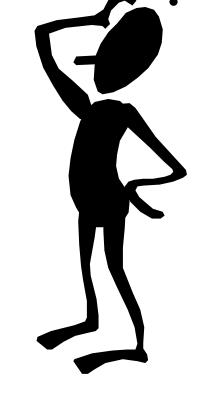


İletişim:

Özgür Yılmazel

Anadolu Üniversitesi Bilgisayar Mühendisliği

www.ozguryilmazel.com ozgur@yilmazel.com Tel: +90-222-335 0580 / 6555



SORULAR?