Bilgiye Erişim Sistemlerinde Arama Kalitesini İyileştirme: Normalleştirme Etkeninin Önemi

Özlem Karagedik1, Ata Önal1

1 Ege Üniversitesi, Bilgisayar Mühendisliği Bölümü, İzmir

ozlemkaragedik@gmail.com, ata.onal@ege.edu.tr

Özet:

Bilgiye erişim sistemlerinde aramanın kalitesi, erişilen ilgili belge sayısı ile doğru orantılıdır. Bu sistemler birbirinden farklı uzunluklarda çok sayıda belge içerebilirler. Belgelere erişirken uzun belgeler kısa belgelerden daha avantajlı hale gelebilirler. Bu durumun önlenmesi ve bütün belgelere aynı şansın verilebilmesi için belge uzunluklarını normalleştirmek gerekir. Çalışmada klasik vektör uzay modelinde kullanılan kosinüs normalleştirmesi ve eksenli benzersiz normalleştirme modeli karşılaştırılmış ve iki model de WikipediaMM 2008 veri kümesi üzerinde denenmiştir. Eksenli benzersiz normalleştirme modelinin daha başarılı olduğu gözlemlenmiştir.

Anahtar Sözcükler: Bilgiye erişim modeli, vektör uzay modeli, WikipediaMM, eksenli benzersiz normalleştirme, kosinüs benzerliği

Improvement of Search Quality In Information Retrieval Systems : Importance of Normalization Factor

In the information retrieval systems, the quality of search is proportional to the number of relevant documents retrieved. These systems can contain numerous documents of different lengths. While retrieving documents, long documents can have advantage over short documents. To prevent this situation and give the same opportunities to all of the documents in the collection, the lengths of all the documents must be normalized. In this work, the cosine normalization and the pivoted unique normalization is compared and tested on WikipediaMM 2008 dataset. Experimentation shows that pivoted unique normalization is more successful than cosine normalization.

Abstract:

Keywords: Information Retrieval, Vector Space Model, WikipediaMM, Pivoted Unique Normalization, Cosine Similarity,

1. Giriş

Bilgiye erişim sistemleri, belge arşivlerinde kullanıcıların isteklerine uygun belgelere, kullanıcıların kolay bir şekilde erişimlerini sağlayan sistemlerdir. Bu sistemlerin temel amacı, kullanıcıların bilgi ihtiyaçlarını karşılamak için, belge arşivlerindeki ilgili (relevant) belgelerin hepsine erişmek, ilgisiz (non-relevant) belgeleri ise çıkartmaktır.

İnternetin yaygınlaşmasıyla daha da büyüyen veri havuzundaki bilginin çıkarılması günümüzün en popüler konularından biri olmuştur. Kullanıcıların arama motorlarında yaptıkları sorgulamalarının sonuçlarının kullanıcıların isteklerine en iyi cevabı verebilmesi de büyük önem kazanmıştır [9]. Bilgiye erişim modeli, kullanıcının ihtiyaç duyduğu belgeye ulaşırken, belge arşivindeki sorgulamasını, belgenin içerdiği kelimeler ile yapmasına olanak tanımaktadır.

Çalışmada, kullanıcının ihtiyacı olan ilgili belgelere en iyi oranda erişerek arama kalitesinin arttırılması amaçlanmıştır. Bu amaçla, vektör uzay modeli ve eksenli benzersiz normalleştirme modeli kullanılarak bu modelinin arama kalitesindeki olumlu etkileri gözlemlenmiştir.

Bilgiye erişim modelinde kullanılan en klasik yöntemlerden biri de vektör uzay modelidir. Salton’un vektör uzay modeli, her bir belgeyi içerisindeki terimlerin ağırlığından oluşan bir vektör olarak tanımlar [1]. Her bir terimin ağırlığını hesaplarken, terimin belgede geçme sayısıyla, bütün belge arşivinde geçme sayısını oranlayarak bir ağırlık elde eder [1]. Bir terimin belgedeki ağırlığı hesaplanırken, uzun belgeler kısa belgelere göre avantajlı duruma geçebilir. Bu yüzden belge uzunluklarını normalleştirmek gerekmektedir [8]. Belge uzunluklarının normalleştirilme gereksinimleri şunlardır:

a. Yüksek terim frekansları: Uzun belgeler, genelde aynı terimi çokça kez tekrar eder. Bu yüzden terim sıklık etkeni uzun belgeler için çok yüksek olur. Bu da belgedeki terimlerinin ağırlığının artmasına; sorgu ve belge benzerliği değerinin yüksek olmasına ve uzun belgelerin kısa belgelere göre daha avantajlı hale gelmesine sebep olur [2].

b. Fazla sayıda terim: Uzun belgeler fazla sayıda farklı terim içerir. Bu da bir sorgu ile belgenin eşleşme sayısını arttırırken aynı zamanda da belge doküman benzerliğini arttırır ve erişimde kısa belgelere göre uzun belgeleri daha şanslı bir konuma getirir [2].

Çalışmada ilk yöntem olan vektör uzay modelindeki normalleştirme için kosinüs benzerliği kullanılmıştır. Belgeler vektör uzunluklarına bölünerek birim vektör haline getirilmişlerdir. Fakat bu durumda belgelerin uzunlukları tamamen göz ardı edilmiştir. Belgedeki normalizasyon faktörü yine terimin ağırlığına bağımlı kalmıştır. Belgeler büyüdükçe kosinüs benzerliğinin erişim performansında zayıf kaldığı gözlemlenmiştir.

Çalışmadaki ikinci yöntem olan eksenli benzersiz normalleştirme modelinin [5], farklı uzunluktaki belgeler için çıkan sorunlara daha etkin bir çözüm getirerek, daha başarılı olduğu gözlemlenmiştir. Bu durumda vektörlerin birim vektör olması gerekmemektedir. Eksenli benzersiz normalleştirme modeli belgelerdeki farklı terim sayısını bir normalleştirme etkeni olarak kullanarak, farklı uzunluklardaki tüm belgelerin erişim aşamasında aynı şansa sahip olmasını sağlamaktadır. Sonuç olarak bu durum da erişim performansı arttırmaktadır.

Çalışmanın ilerleyen bölümleri şu şekilde düzenlenmiştir: İkinci bölümde sorguların belgeler ile karşılaştırılması, birbirlerine benzerliklerinin bulunması ve belgelerin sıralanması için çalışmada kullanılan yöntemler üzerinde durulacaktır. Üçüncü bölümde, üzerinde çalışılan veri kümesi anlatılacaktır. Dördüncü bölümde, ikinci bölümde anlatılan yöntemlerle veri kümesi üzerinde yapılan deneysel çalışmalara değinilecektir. Son bölümde ise çalışmanın şimdiki durumu ve çalışmanın devamında arama kalitesine arttırabilmek için uygulanacak olan diğer yöntemler üzerinde durularak makale sonlandırılacaktır.

2. Yöntem

Çalışmada, sorguların belgeler üzerinde aranması işlemlerine başlamadan önce hem belgeler hem de sorgular bir takım ön işlemlerden geçirilirler. Bunlardan ilki tek harfli veya iki harfli gibi çok kısa kelimelerin silinmesidir. Daha sonra atılacak kelimeler listesindeki (stop word list) kelimelerin atılması aşaması gelir (stop word elimination). Atılacak kelimeler listesi, içerisinde hemen her belgede sık sık geçebilecek veya erişim değeri olmayan kelimelerden oluşur. Üçüncü aşama ise kelimenin anlamsal olarak köküne inme aşamasıdır. Bu aşamada kelimelerdeki ekler atılarak kelime kök haliyle bırakılır.

Belgeler ve sorgular, ön işlemden geçirildikten sonra bilgiye erişim sisteminde işlenmek, ve sorgulara ait ilgili belgelerin bulunması için bir sonraki aşamaya geçmeye hazır hale getirilir.



Şekil 1. Genel Mimari

2.2. Vektör Uzay Modeli

Vektör Uzay Modeli, bilgiye erişim alanında sıklıkla kullanılan geleneksel bir yöntemdir. Vektör Uzay Modelinde, belgelerin ve bu belgeler üzerinde arama yapılan sorguların her biri birer vektör ile ifade edilir [6]. Her bir belge, anlamlı kelimelerden oluşan ve kelimelerin her birinin ağırlığı olan birer terim vektörü şekline getirilir. Belge vektörleri bir araya getirilerek bütün belgeleri içeren bir matris oluşturulur. Bu matris, belgelerden ve belgelerdeki terimlerden oluştuğu için belge terim matrisi ( matrisi) adını alır [1]. belgeden ve terimden oluşan bir belge terim matrisi örneği denklem 1’de verilmiştir. Matrisin her bir satırı bir belgeyi vektör olarak gösterirken, sütunları da terim vektörleri olarak adlandırılır[6]. Her bir hücredeki ise terimin belgedeki ağırlığıdır. Aşağıda bir belge terim matrisi örneği gösterilmiştir.

 (1)

Terim ağırlığı hesaplamanın bir çok değişik yolu vardır. Terim sayısı modeli bunlardan biridir. Bu modelde teriminin belgedeki ağırlığı, belge içerisindeki geçiş sayısına eşittir. Vektör Uzay modeli için en çok kullanılan model ise, term sıklığı - devrik belge sıklığı ağırlık modelidir [6].

**Terim Sıklığı (Term Frequency) - Devrik Belge Sıklığı (Inverse Document Frequency ) (Ts-Dbs)**

Ts-Dbs ağırlık modeli, bir terimin belirli bir belge için, belgenin bulunduğu belge havuzu içerisinde ne kadar önemli olduğunu ölçmek için kullanılır[6]. Bu modele göre, terimin önemi, belge içerisinde o terimin geçme sayısıyla doğru orantılıyken; bütün belge havuzu içerisinde o terimin geçme sıklığıyla ters orantılıdır. Denklem 2’de de belirtildiği gibi terimin ağırlığı şu şekilde hesaplanır [6].

 (2)

Burada terimin belge içerisindeki geçme sıklığını, belge havuzundaki döküman sayısını, ise i terimini içeren belge sayısını ifade etmektedir. dfi /D oranı, belge havuzunda i terimini içeren belgeyi seçme olasılığıdır. Böylece devrik belge sıklığını ifade etmektedir. Devrik belge sıklığı, i. teriminin belge havuzu için global önemini belirtmektedir [6].

**Kosinüs Benzerliği**

Kosinüs benzerliği iki vektör arasındaki kosinüs uzaklığını bularak bu vektörlerin birbirleriyle benzerliklerini ölçmek için kullanılmaktadır [1]. Kosinüs benzerliği ile bir belge vektörü ve bu belge vektörüyle benzerliğini bulacağımız bir sorgu vektörü karşılaştırılarak aralarındaki ilişki bulunur. Kosinüs benzerliğinde belge ve sorgular kendi vektör uzunluklarına bölünerek birim vektör haline getirilir [1]. Belge ve sorgunun kosinüs benzerliği 0 ile 1 arasında bir değer alır. Kosinüs benzerliği değerinin 1’e yaklaşması belge ve sorgunun birbirleriyle olan benzerliği daha fazla olduğu anlamına gelir. Benzerlik fonksiyonu aşağıda görüldüğü gibidir.

 (3)

 (4)

Burada , Q sorgusuyla ve i belgelerinin nokta çarpımları ise Q sorgusu ve i belge vektörlerinin uzunluklarının çarpımıdır. Kosinüs benzerliğinde, vektör uzunlukları normalleştirme etkeni olarak kullanılır. değeri kosinüs normalleştirme etkeni olarak bilinir [6]. Kosinüs normalleştirmesi, uzun belgeler için fazla terim sayısı ve yüksek terim sıklığı probleminin üstesinden gelmeye çalışsa da; gerçek belge uzunluklarını dikkate almadığı için, belgelerin uzunluğu büyüdükçe erişim başarısı düşer [4].



**Şekil 2.** Kosinüs Benzerliği

**2.3. Eksenli Benzersiz Normalleştirme Modeli (EBNM)**

Eksenli benzersiz normalleştirme modeli klasik vektör uzay modelinin değiştirilmiş bir versiyonudur. Terim ağırlığı hesaplanırken, terim ve belge frekanslarından farklı olarak bir normalleştirme etkeni de denkleme eklenir. Bu modele göre bir terimin belge içerisinde ağırılığı Denklem 5 ile hesaplanmaktadır [5].

 (5)

burada terimin belgede geçme sayısını, belge içerisindeki tüm terimler için ifadesinin toplamını, *N* koleksiyondaki toplam belge sayısını, *nf* j. terimi içeren belge sayısını ve *U*’da belgedeki benzersiz terim sayısını gösterir [5]. Benzersiz ifadesi, belgenin uzunluğunun belgedeki benzersiz kelime sayısıyla ölçülmesidir [5].

Benzersiz terim sayısı(U), denklemin ikinci çarpanı olan normalleştirme etkeninde kullanılır. Normalleştirme etkeni uzun belgelerin fazla sayıda farklı terim içermesi ve bunun uzun belgelerin ağırlıklarını yükseltmesi problemine çözüm getirir [2][4]. Böylece normalleştirme etkeni kullanılarak, belgenin uzunluğunun ortalama uzunluktan daha kısa olduğu durumda ağırlığı arttırılmakta, belge ortalama uzunluktayken ağırlığı aynı bırakılmakta, belge ortalama uzunluktan daha uzun ise ağırlığı azaltılmaktır [2]. 0.0118 değeri ise denklemde kullanılan eksen değeridir.

Denklemin ilk kısmında, terimin belgede geçme sayısının, değerine bölünmesiyle, Bölüm 1’ de bahsettiğimiz, uzun belgelerde aynı sayıda terimin çokça geçmesi ile terim sıklıklarının yüksek olması problemine de çözüm getirilmiş olunur [2].

Belge ve sorgular arasında benzerlik sıralaması değeri ise denklem 5’ de bulduğumuz değerinin, terimin sorgudaki frekansıyla çarpılmasıyla bulunur. Bu model için sıra değeri denklem 6’daki gibi hesaplanmıştır [5].

|  |  |
| --- | --- |
|  | (6) |

Burada n sorgudaki terim sayısını, terimin belgedeki ağırlığını ve terimin sorguda kaç kere geçtiğini ifade eder [5].

**3. Veri Kümesi**

WikipediaMM, görsel bilgiye erişim için Wikipedia resimlerinin koleksiyonundan oluşan bir test ortamı sunar [3]. Veri kümesi olarak Wikipedia’nın kullanılmasının amacı, kullanıcının arama yapabileceği, web’e benzer nitelikte daha büyük ölçekli ve heterojen içerikli bir koleksiyonu araştırmaktır. Veri kümesi, çeşitli konularda 151,519 resim içerir [3]. Bu resimlerin her biri yapısal olmayan gürültülü metinsel açıklamalar içerir. Çalışmada WikipediaMM 2008 veri kümesi kullanılmıştır. Bu veri kümesinde resimler üzerinde aranabilecek 75 metinsel sorgu bulunur [3].

**4. Deneysel Sonuçlar**

Bu çalışmanın amacı WikipediaMM gibi büyük ve heterojen olan bir veri kümesinde en iyi sonucu elde etmektir. Bir başka ifadeyle ilgili belgelerin kullanıcıya ilk sıralarda gösterildiği en iyi yöntemi bulmaya çalışmaktır. Bu çalışmada farklı uzunluklardaki belgelere aynı şansın verilmesi gerektiğinden bahsedilmiştir. Bunun için de klasik vektör uzay modelinde kullanılan kosinüs normalleştirmesi ve eksenli benzersiz normalleştirme modeli karşılaştırılmıştır.

Bilgiye erişim sistemlerini değerlendirmek ve sistemin kalitesini belirleyebilmek için çeşitli değerlendirme ölçüleri vardır.

İlgililik (relevance) kavramı sistemden dönen sonuçların kullanıcının ihtiyaçlarına ne kadar iyi örtüştüğüdür. Sistemin kalitesine belirleyen ölçütlerden en önemli iki tanesi,

hassasiyet (H)(Precision) ve anmadır (A)(Recall) [7] . Hassasiyet ve anma değerleri aşağıdaki gibi hesaplanır. Hassasiyet erişilen ilgili belge sayısının, erişilen bütün belgelere oranıyla bulunurken; anma, erişilen ilgili belge sayısının, bütün ilgili belge sayısına oranıdır [7].

Standart bir değerlendirme stratejisi, ortalama hassasiyet ve anma değerlerini bulmaktır. Bu grafik farklı sorguların veri kümesi üzerinde çalıştırılıp bu sorguların belirli anma seviyelerinde ortalama hassasiyet değerlerinin hesaplanmalarıyla bulunur [2]. Değerlendirme ölçüsü tek bir değer şeklinde de elde edilebilir. Buna ortalama hassasiyet adı verilir(MAP) [2]. Sorgu bazında ilgili doküman seviyelerindeki hassasiyet değerlerinin ortalaması alındıktan sonra, bütün sorgular için bu değerler toplanarak ortalaması alınır. Böylece sistemi ölçmek için tek bir değer olan ortalama hassasiyet bulunmuş olunmaktadır.

Bunun içinde WikipediaMM veri kümesi üzerinde çalıştırılan her iki yöntem için de, 75 sorgu çalıştırılarak, her sorgu için sıralama değeri en yüksek olan ilk 1000 belge alınmıştır. Şekil 2.de 1 numaralı yöntem olan, klasik vektör uzay modeli ve kosinüs benzerliğini kullanan sistemin sonuçlarını, iki numaralı yöntem olan eksenli benzersiz normalleştirme modelinin sonuçları görülmektedir. Tablo 1’de ise bir başka değerlendirme kriteri olan ortalama hassasiyet değerlerini, ilk 5 belge ve ilk 10 belgedeki hassasiyet değerlerini, erişilen belge sayısını, erişilen ilgili belge sayısını, ilgili belge sayısını görmekteyiz.

Bu sonuçları inceleyecek olursak eksenli benzersiz normalleştirmenin ortalama anma ve hassasiyet grafiği değerlerinde üstün olduğu görülebilir. Aynı şekilde Tablo 1’de de görüleceği gibi, ilk yöntem toplamda 2369 ilgili belgeye erişirken ikinci yöntem ile bu sayı 3184’e çıkmıştır. İkinci yöntemin ortalama anma değeri 0,2549 iken, ilk yöntemde 0,1394 olarak gözlemlenmiştir. Bütün sonuçlara bakılarak ikinci yöntem olan eksenli benzersiz normalleştirme modelindeki belge uzunluğunu normalleştirme yönteminin, klasik vektör uzay modelinden daha başarılı bir şekilde belgelere eriştiğini söyleyebiliriz.

**5. Sonuç**

Bilgiye erişim sistemlerinde, kullanıcının ihtiyaçları doğrultusunda, ilgili belgelere ulaşmasının önemi büyüktür. Bu da arama kalitesinin arttırılması bir başka ifadeyle aramanın yüksek hassasiyet değerlerine sahip olması gerektiği anlamına gelir.

WikipediaMM farklı uzunluklarda ve heterojen içerikli çok fazla sayıda belgenin bulunduğu bir belge arşividir. Bu arşiv üzerinde yapılan çalışmanın deneysel sonuçlarına baktığımızda, klasik vektör uzay modelinin değiştirilmiş bir versiyonu olan Eksenli benzersiz normalleştirme modelinin daha başarılı olduğu gözlemlenmiştir. Bu model belgelerin uzunluklarını dikkate alan bir normalleştirme etkeni kullanarak, çok büyük boyuttaki belgelerin ağırlıklarını azaltılmış, küçük boyuttaki belgelerin ağırlıklarını arttırmış ve bütün belgelerdeki ağırlıkları ortalama boyuttaki belgelerle ile aynı seviyeye getirmiştir. Bu durum da eksenli benzersiz normalleştirmeyi bilgiye erişimde güçlü ve etkili bir yöntem haline getirmiştir.

Çalışmanın bundan sonraki ilerleyişi kullanıcının sorgularını sistemin otomatik olarak genişletmesi ile arama kalitesinin arttırılması üzerine olacaktır. Genişletme ile, sorgudaki kelimeler ile benzer anlamlı yeni kelimeler, sorguya eklenecektir. Böylelikle sorgunun içeriği zenginleştirilerek, sorgunun ilk halindeki kelimeyi içermeyen ama sorgu genişletilince bulunabilen ilgili belgelere de erişimi sağlamak amaçlanacaktır.

**6. Referanslar**

[1] Salton G, Wong A, Yang CS. A Vector Space Model For Automatic Indexing. *Communications of ACM*. 1975. 18 (11): 613-620

[2] Singhal A, Buckley C and Mitra M. Pivoted Document Length Normalization. Proceedings of SIGIR. 1996. p. 21-29

[3] Theodora Tsikrika and Jana Kludas. Overview of the wikipediaMM task at ImageCLEF 2008. In Evaluating Systems for Multilingual and Multimodal Information Access, Proceedings of the 9th Workshop of the Cross-Language Evaluation Forum, Lecture Notes in Computer Science, vol. 5709, pp. 539-550, Springer 2009.

[4] Manning D. Chirstopher, Raghavan Prabhakar and Schütze Hinrich. An Introduction to Information Retrieval, Cambridge University Press, 2009.

[5] E. Garcia. Implementation and application of term weights in mysql environment, 10 2006.

[6] E. Garcia. The Classic Vector Space Model,10,2006.

[7] Singhal A, "Modern Information Retrieval: A Brief Overview,2006.

[8] April Kontostathis and Scott Kulp, The Effect of Normalization when Recall Really Matters,2008.

**[**9] Ricardo Baeza-Yates,Berthier Ribeiro-Neto,Modern Information Retrieval,1999.

Tablo 1. Wiki 2008 veri kümesinde performans sonuçları

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **No** | **Yöntem** | **Ort. Hassasiyet** | **İlk 5’de hassasiyet** | **İlk 10’da hassasiyet** | **Erişilen**  | **İlgili Erişilen** | **İlgili** |
| 1 | Vektör Uzay Modeli-Kosinüs Norm. | 0.1394 | 0.2427 | 0.2013 | 37500 | 2369 | 5593 |
| 2 | Eksenli Benzersiz Normalleştirme | **0.2549** | **0.4453** | 0.3693 | 65888 | 3184 | 5593 |



Şekil 2.Wiki 2008 H/A Grafiği