Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü
Bilgi ve Belge Yönetimi Anabilim Dalı

SINIRLI ALANLARDA KONU TESPİT VE TAKİBİ İÇİN GENİŞLETİLMİŞ BİR MİMARİ YAPI ÖNERİSİ

Güven KÖSE

Doktora Tezi

Ankara, 2014

SINIRLI ALANLARDA KONU TESPİT VE TAKİBİ İÇİN GENİŞLETİLMİŞ BİR MİMARİ YAPI ÖNERİSİ

Güven KÖSE

Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü
Bilgi ve Belge Yönetimi Anabilim Dalı

Doktora Tezi

Ankara, 2014

KABUL VE ONAY

Güven KÖSE tarafından hazırlanan "Sınırlı Alanlarda Konu Tespit Ve Takibi İçin Genişletilmiş Bir Mimari Yapı Önerisi" başlıklı bu çalışma, 5 Haziran 2014 tarihinde yapılan savunma sınavı sonucunda başarılı bulunarak jürimiz tarafından Doktora Tezi olarak kabul edilmiştir.

Prof. Dr. Serap Kurbanoğlu (Başkan)

Prof. Dr. Yaşar Tonta (Danışman)

Doç. Dr. Hasan Oğul

Doç. Dr. İrem Soydal

Yrd. Doç. Dr. Erhan Mengüşoğlu

Yukarıdaki imzaların adı geçen öğretim üyelerine ait olduğunu onaylarım.

Prof. Dr. Yusuf Çelik
Enstitü Müdürü

BİLDİRİM

Hazırladığım tezin/raporun tamamen kendi çalışmam olduğunu ve her alıntıya kaynak gösterdiğimi taahhüt eder, tezimin/raporumun kâğıt ve elektronik kopyalarının Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü arşivlerinde aşağıda belirttiğim koşullarda saklanmasına izin verdiğimi onaylarım:

■ Tezimin/Raporumun tamamı her yerden erişime açılabilir.

☐ Tezim/Raporum sadece Hacettepe Üniversitesi yerleşkelerinden erişime açılabilir.

☐ Tezimin/Raporumun 2 yıl süreyle erişime açılmasını istemiyorum. Bu sürenin sonunda uzatma için başvuruda bulunmadığım takdirde, tezimin/raporumun tamamı her yerden erişime açılabilir.

05.06.2014

Güven KÖSE

TEŞEKKÜR

Tez konusunun şekillendirilmesinden başlayarak başarıyla sonlandırılmasına kadar tüm aşamalarda, değerli katkılarını ve desteğini esirgemeyen tez hocam Prof. Dr. Yaşar Tonta'ya, tez çalışmaları esnasında katkılarını esirgemeyerek çalışmaların daha verimli yürütülmesine destek olan tez izleme komitesindeki değerleri hocalarım Prof. Dr. Serap Kurbanoğlu ve Doç. Dr. Hasan Oğul'a özel olarak teşekkür ederim.

Tez çalışması esnasında oluşturdukları derlemi kullanımımıza açma nezaketi gösteren Prof. Dr. Fazlı Can ve Bilkent Üniversitesi Bilgi Erişim Grubu'nun değerleri üyelerine teşekkür ederim.

Ayrıca hem tezin yazımı aşamasında hem de hayatımın tüm aşamalarında güçlü desteklerini eksik etmeyen sevgili eşim Derya Selçuk Köse'ye ve değerli aileme çok teşekkür ederim.

ÖZET

KÖSE, Güven. Sınırlı Alanlarda Konu Algılama Ve İzleme İçin Genişletilmiş Bir Mimari Yapı Önerisi, Doktora Tezi, Ankara, 2014.

Internet üzerindeki bilginin devasa boyutlara ulaşması ile birlikte bu mecra bilgi arayan kullanıcıların birinci tercihi haline gelmiştir. Kullanıcıların Internet üzerindeki bilgiye karşı olan bu yoğun ilgisi hem arama motorlarının hem de bilgi erişim sistemlerinin önemini bir kat daha artırmıştır. Internet üzerinde sınırlı sayıda kelime ile bilgi arayan kullancılar, arama motorlarını yoğun olarak kullanırken, daha özel ve derinlemesine bilgi ihtiyacı olan kullanıcılar, özelleşmiş bilgi erişim sistemlerini kullanmaktadırlar. Bu kapsamda özelleşmiş bilgi erişim sistemleri ile ilgili çalışmalar son yıllarda yoğun olarak haber algılama ve izleme sistemleri olarak da tanımlanabilecek "Konu Algılama ve İzleme" programı üzerinde yoğunlaşmıştır. Bu programdaki çalışmaları geleneksel bilgi erişim sistemlerinden ayıran en önemli unsur, bilgi erişim sistemlerinde kullanılan sorgu-belge eşleşmelerinin yerini belge-belge eşleşmelerinin almış olmasıdır. Buna ek olarak, sisteme ulaşan bağımsız iki haberin aynı konuda olup olmadığını anlamaya çalışan "hikâye bağlantı algılama" ve önceden belirlenmiş bir konuda yeni çıkan haberleri yakalamayı hedefleyen "konu izleme" görevleri bu programın en önemli parçaları olarak tanımlanmıştır.

Bu çalışma kapsamında, hikâye bağlantı algılama ve konu izleme görevlerinin gerçekleştirilmesinde farklı erişim fonksiyonu ve belge gösterim tekniklerinin başarım üzerindeki etkileri araştırılmıştır. Bu bağlamda, hikâye bağlantı algılama görevinin başarımını test etmek için vektör uzayı modeli ve ilgi modeli erişim fonksiyonu olarak kullanılmıştır. Buna ek olarak, belge gösterim tekniği olan *tf.idf* değerlerinden en yüksek olan terimler seçilerek bu terimlerle başarım testleri tekrarlanmış ve her bir yöntem için en uygun terim sayıları belirlenmiştir. Ayrıca, konu izleme görevi ile ilgili olarak uygun eşik değerinin seçilmesinin ve erişim fonksiyonu olarak vektör uzayı, ilgi modeli ve k-ortalamalar yöntemlerinin kullanılmasının başarım üzerindeki etkileri araştırılmıştır.

Gerek hikâye bağlantı algılama gerekse konu izleme ile ilgili başarım testleri daha önce benzer akademik çalışmalarda kullanılmış olan BilCol-2005 Türkçe haber derlemi kullanılarak gerçekleştirilmiştir. Bu derlem üzerinde gerçekleştirilen başarım testlerinin f-ölçü sonuçlarına göre, hikâye bağlantı algılama görevinde vektör uzayı modelinin ilgi modeline göre çok daha yüksek bir başarıma sahip olduğu belirlenmiştir. Ayrıca, belge gösteriminde vektör uzayı modelinde 30 terim, ilgi modelinde ise 4 terim için en yüksek f-ölçü değerlerine ulaşılmıştır. Konu izleme görevinde, anma ve duyarlığın en yüksek olduğu noktadaki değerin eşik değeri olarak seçilmesinin en başarılı yöntem olduğu belirlenmiştir. Bunun yanında k-ortalamalar yönteminin konu izleme görevinde en başarılı yöntem olduğu tespit edilmiştir.

Ayrıca bu çalışma kapsamında, hikâye bağlantı algılama ve konu izleme görevleri için gerçekleştirilen başarım testlerinden elde edilen sonuçlar ışığında, elimizde eğitim belgelerinin bulunmadığı durumlar için Türkçe bir konu izleme sistemi önerilmiştir. Bu sistemde konu modellerini oluşturmak ve zenginleştirmek için vektör uzayı ve ilgi modellerinin AND birleşimlerinin kullanılması önerilmektedir. Ayrıca sisteme yeni ulaşan haberlerin konu modeli ile ilgili olup olmadığının tespit edilebilmesi için k-ortalamalar yöntemi kullanılmalıdır. Önerilen bu mimari yapı ile Türkçe için etkin bir izleme sistemi oluşturulabileceği düşünülmektedir.

Anahtar Sözcükler

Konu algılama ve izleme, hikâye bağlantı algılama, konu takibi, bilgi erişim sistemleri, Türkçe konu takip sistemi.

ABSTRACT

KÖSE, Güven. A Proposal of an Extended Architecture for Topic Detection and Tracking in Limited Domains. Ph. D. Dissertation, Ankara, 2014.

As the rate of growth of information on the Internet is enormous, the need for retrieving the right information has become one of the most important things for the users. Users that need specific and deep information aim to use advanced information retrieval technologies, while other users use the search engines with restricted keywords. In this context, "Topic Detection and Tracking" program, which can be defined as news detection and tracking systems, has become one of the most important attraction centers of research. The most important factor of this system that differs from other traditional information retrieval systems is that this system uses document-document matching instead of query-document matching. In addition to this, The "Story Link Detection" detects two similar stories within the system whether they have the same subject or not while the "Topic Tracking" has the target of catching the news updates for a predefined subject. These two properties are considered as the two most important parts of the system.

This study investigates the effects of different retrieval functions and document representation techniques on performance in carrying out the tasks of story link detection and topic tracking. In this context, vector space and relevance models were used as retrieval functions. In addition, terms that scored the highest *tf.idf* values have been selected for document representation, performance tests have been repeated with these terms, and the most appropriate terms for each method have been identified. Moreover, the effects of choosing the appropriate threshold values for topic tracking on performance along with vector space, relevance model and k-means methods as retrieval functions have been examined.

Both story link detection and topic tracking performance tests have been fulfilled by the use of BilCol-2005 Turkish news corpus used in similar studies. Vector space model scored higher f-measure values on this corpus than that of relevance model in performance tests for story link detection tasks. The highest f-measure values for

document representation were obtained for 30 and 4 terms in vector space and relevance models, respectively. Choosing the threshold value where precision and recall values were the highest turned out to be the most successful method for topic tracking along with k-means method.

In the light of the findings obtained from performance tests carried out for story link detection and topic tracking tasks, a topic tracking system for Turkish corpora where no training documets exist has been proposed. The AND combination of the vector space and the relevance models should be used in order to create and enrich topical models. Also, k-means method should be used to determine if incoming news items are related with the topical model. We think the proposed architecture can help to build an effective topic tracking system for Turkish.

Keywords

Topic detection and tracking, story link detection, topic tracking, information retrieval systems, topic tracking systems in Turkish.

İÇİNDEKİLER

KISALTMALAR

TDT	Topic Detection And Tracking
TF	Term Frequency
IDF	Inverse Document Frequency
VUM	Vektör Uzayı Modeli
İM	İlgi Modeli
BES	Bilgi Erişim Sistemi
SLD	Story Link Detection
NED	New Event Detection
KNN	K-Nearest Neighbor
LSI	Latent Semantic Indexing
SVM	Support Vector Machines
TS	Terim Sayısı

TABLOLAR

ŞEKİLLER

ÖNSÖZ

Geleneksel bilgi erişim sistemlerinin devamı olarak düşünülen konu algılama ve izleme programında tanımlanmış görevlerle ilgili olarak Türkçe çalışmaların sayısının sınırlı olması, bu çalışmanın en önemli motivasyon kaynağını oluşturmuştur. Bu çalışmadan elde edilecek sonuçların bundan sonra gerçekleştirilecek olan akademik çalışmalara bir temel oluşturması beklenmektedir.

Bu tez çalışması, alanda gerçekleştirilen sınırlı çalışmalardan birisi olan *"Türkçe Haber Benzerliklerinin Belirlenmesinde Varlık İsimlerinin Hikaye Bağlantı Algılama Görevinin Başarımına Etkisi"* isimli TÜBITAK projesi (TÜBİTAK Sosyal Bilimler Araştırma Grubu - Proje No: SOBAG 111K030) tarafından desteklenmiştir.

Projede ağırlıklı olarak iki haberin aynı konuda olup olmadığına karar verirken haberlerde geçen varlık isimlerinin başarım üzerindeki etkileri araştırılmıştır. Bununla birlikte bu tez çalışmasında hikaye bağlantı algılama ve konu izleme görevlerinin gerçekleştirilmesi ile ilgili sonuçlar projede bir başlangıç noktası olarak kabul edilmiştir. Başarım testleri Bilkent Üniversitesi Bilgi Erişim Grubu tarafından hazırlanmış olan BilCol-2005 haber derlemi kullanılarak gerçekleştirilmiştir. Derlem üzerinde ayrıntılı olarak bir etiketleme çalışması gerçekleştirilmiş ve haberler kim (who), ne (what), ne zaman (when) ve nerede (where) sorularına yanıt verecek şekilde yeniden düzenlenmiştir.

Proje Ocak 2014 tarihinde başarıyla sonlanmış ve proje sonuç raporu yayımlanmıştır. Proje kapsamında elde edilen sonuçlar, 4-6 Eylül 2013 tarihleri arasında Hacettepe Üniversitesi ve Limerick Teknoloji Enstitüsü tarafından düzenlenen 4th International Symposium on Information Management in a Changing World adlı toplantıda “Supervised news classification based on a large-scale news corpus” (Köse ve Ahmadlouei, 2013) ve 17-20 Kasım 2013 tarihleri arasında IEEE tarafından düzenlenen International Conference on Web Intelligence toplantısında ise “Story link detection in Turkish Corpus” (Köse, Tonta, Ahmadlouei ve Polatkan, 2013) isimli çalışmalar ile tarafımızdan duyurulmuştur.

Proje kapsamında, "Sınırlı Alanlarda Konu Tespit ve Takibi İçin Genişletilmiş Bir Mimari Yapı Önerisi" isimli bu doktora tezinin yanı sıra "Türkçe Haber Benzerliklerinin Belirlenmesinde Varlık İsimlerinin Etkisi" (Hamid Ahmadlouei) ve "Haber Metinlerinin Kategorizasyonunda Varlık İsimleri ve Konu Başlıkları İlişkisi" (İpek Şencan) isimli iki yüksek lisans tezi başlatılmış ve çalışmalardan bir tanesi tamamlanmış diğeri de sunum aşamasına gelmiştir.

Bu bağlamda, bu tez kapsamında ilerleyen bölümlerde paylaşılan içerik ve sonuçların belli bir kısmı ilgili projenin yukarıda anılan sonuç raporunda da kısmen ya da tamamen yer almaktadır. Proje sonuç raporunda yer alan analizler tarafımızdan gerçekleştirilmiş ve rapor edilmiştir. Ancak proje sonuç raporu bu doktora tezinden önce TÜBİTAK'a sunulmuştur. Bu nedenle proje sonuç raporunda kullanılan ve fakat bu tezin esas araştırma konusunu oluşturan bulgular ve yorumlar için proje sonuç raporu kaynak olarak gösterilmemiştir. Rapordan aynen alınan tablolar için ise proje sonuç raporundaki sayfa numaraları dipnotlarda belirtilmiştir.

"Türkçe Haber Benzerliklerinin Belirlenmesinde Varlık İsimlerinin Hikaye Bağlantı Algılama Görevinin Başarımına Etkisi" isimli projenin yürütülmesinde ve elde edilen sonuçların tez çalışmasını desteklemesinde önemli katkıları olan proje yürütücüsü Doç. Dr. İrem Soydal'a ve proje araştırmacısı Doç. Dr. Umut Al' a teşekkürü bir borç bilirim.

1. BÖLÜM

GİRİŞ

1.1. KONUNUN ÖNEMİ VE AMAÇ

Internet'in hızla gelişmesi ve yaygınlaşması, kullanıcı ilgisinin geleneksel bilgi erişim sistemlerinden (BES), web üzerinde arama yapan arama motorlarına (search engines) doğru kaymasına neden olmuştur (Gaines, Chen ve Shaw, 1997; Lawrence ve Giles, 1998). Internet üzerinde bilgi arayan kullanıcıların pek çoğu, popüler arama motorlarını gerekli kaynaklara ulaşmak için tek yol olarak görmektedirler. Bu nedenle, tercih edilen arama motorlarının, kullanıcıların bu taleplerini karşılayabilmek için gerekli yeterliliğe sahip olmaları son derece önemlidir. Kullanıcı açısından bakıldığında, ideal bir arama motorunun Internet üzerindeki bütün bilgi ya da belgeye erişim sağlaması, arama sonuçlarını çok hızlı bir biçimde sunması, eriştiği bütün sonuçların güncel olması, arama sonucu erişilen bütün belgelerin sorgu ile ilgili olması, sorgu ile ilgili bütün belgelere erişilebilmesi ve sorgu ile en ilgili belgelerin erişim çıktısında en üstte, daha az ilgililerin ise daha alt sıralarda yer alması beklenir. Ancak, Internet üzerindeki bilginin devasa boyutlara ulaşmasıyla birlikte, kullanıcıların bilgi gereksinimlerini karşılaması beklenen arama motorları yetersiz kalmaya başlamıştır (Balabantaray, Swain ve Sahoo, 2013).

Bilgi ihtiyacının olduğu alana yabancı olan ya da başlangıç seviyesinde bilgiye sahip olan kullanıcıların, uygun sorguları oluşturamaması, sorgu terimi olarak yazılan birkaç kelime ile arama motorlarından harikalar yaratması beklenmesi, sorgu terimi olarak seçilen kelimelerin bağlamlarının çok net belli olmaması ve arama motorlarında zamana bağlı olarak konu takibi yapılamaması, en önemli bilgi erişim sorunları olarak görülmektedir. Bilgi erişim konusundaki sorunlar incelendiğinde, bunların önemli bir kısmının kullanıcı davranışlarından, kalanlarının da bilgi erişim sistemlerinin mantıksal organizasyonundan kaynaklandığı söylenebilir.

Kullanıcı davranışlarının, bilgi erişim problemi üzerindeki etkilerinin araştırıldığı bir çalışmada, arama terimlerini kullanarak sorgular oluşturan kullanıcıların davranışlarını inceleyen araştırmacılar (Jansen, Spink ve Saracevic, 2000), sorgunun konusunun ve kullanıcının arama konusundaki tecrübesinin, oluşturulan sorguları doğrudan etkilediğini göstermişlerdir. Jansen ve diğerleri (2000), Excite arama motoru üzerinde gerçekleştirdikleri çalışmada; kullanıcılar tarafından gerçekleştirilen sorgularda kullanılan terim sayılarını incelemiş ve sorguların yaklaşık %80'inde 3 ya da daha az terim kullanıldığını, sorgu başına düşen ortalama terim sayısının da 2,21 olduğunu saptamışlardır. Diğer taraftan, aynı çalışmada, kullanılan sorgu terimlerinin dağılımları araştırılmış ve çok sayıda terimin, kullanıcı sorgularında az sayıda geçtiği, az sayıda terimin ise sorgularda sıkça kullanıldığı tespit edilmiştir. Bu sonuçlara paralel olarak diğer bir çalışmada; pek çok kullanıcının az sayıda arama, az sayıda kullanıcının da çok sayıda arama gerçekleştirdiği gösterilmiştir (Spink, Wolfram, Jansen ve Saracevic, 2001). Bir başka çalışmada ise; kullanıcıların birden fazla terim kullanarak gerçekleştirdikleri sorgularda, terimlerin birlikte kullanılma sıklıklarının arama davranışları için önemli ipuçları verebileceği gösterilmiştir (Wolfram, 1999). 2000'li yıllarda gerçekleştirilen bu çalışmalara karşılık, daha yeni araştırmalar, kullanıcı davranışlarının günümüzde değişmeye başladığını göstermektedir. Hearst (2011) tarafından gerçekleştirilen bir çalışmada; kullanıcıların bilgi gereksinimlerini elle yazmak yerine mikrofona söylemeyi, metinleri okumak yerine videolarını izlemeyi ve anahtar kelimeler yerine bütün cümleyi girmeyi tercih ettikleri belirtilmektedir. Yine benzer çalışmalarda; geçmiş yıllarda 2 civarında olan ortalama arama terimi sayısının, yıllar içerisinde artış eğiliminde olduğuna vurgu yapılmaktadır (Shah, 2010; Han, Jeong ve Wolfram, 2014).

Bilgiye erişim konusunda, kullanıcı davranışlarından kaynaklanan problemlerin çözümüne yönelik olarak, kullanıcılara rehberlik edebilecek yardımcı kaynaklar geliştirilmeye çalışılmıştır. Alana özel başlıkların (subject headings) yaratılması, kavramsal sözlüklerin (thesauri) ya da terim sözlüklerinin (terms dictionary) oluşturulması, anlamsal ilişkileri yakalayacak ontoloji destekli çözümlerin üretilmesi, varlık isimlerinin belge gösteriminde (document representation) kullanılması ve bibliyometrik analiz yöntemlerinin kullanılması, bunlardan bazılarıdır (Chen, Yim, Fye

ve Schatz, 1995; Castells, Fernandez ve Vallet, 2007; Nowell ve diğerleri, 1996; Soydal ve Al, 2014).

Diğer taraftan, bilgi erişimin mantıksal organizasyonundan kaynaklanan erişim problemlerini (Maron, 1984) daha iyi anlayabilmek için, bu yapıya daha yakından bakmak gereklidir. Bu mantıksal organizasyon içerisinde, bir tarafta sistem tarafından derlem içerisinden çekilerek dizinlenen belgeler bulunurken, diğer tarafta bilgi ihtiyacını karşılamayı amaçlayan kullanıcılar ve bu kullanıcıların bilgi ihtiyaçlarını ifade ettikleri sorgu cümleleri bulunmaktadır. Bu sistem içerisindeki en kritik bileşen, sorgu cümleleri ile dizin terimleri arasındaki çakışmalara göre, sorgu ve belgeler arasındaki benzerlik değerlerini belirleyen erişim fonksiyonudur. Erişim fonksiyonu, doğası gereği, yalnızca kullanıcıların sorgu cümlelerinde geçen ve dizin terimleri ile kesişen belgelere erişim sağlar. Tam da bu noktada, geleneksel bilgi erişim sistemlerinin en büyük problemi olan, dizin terimleri ile kullanıcı sorgularının kesişmemesi durumu ortaya çıkar. Bu sorun, bilgi ihtiyacını karşılamayı amaçlayan kullanıcı açısından bakıldığında, aradığı bilgiyi bulamama anlamına gelmektedir.

Bu problemin çözümüne yönelik olarak kullanıcı tarafında sorgu cümlelerinin genişletilerek sorgu terimlerinin dizin terimleri ile çakışma olasılıklarının artırılması amaçlanır. Erişim fonksiyonu tarafında ise farklı yöntemler ve bu yöntemlerin mantıksal birleşimleri kullanılarak erişim etkinliğinin artırılması hedeflenir.

Arama motorlarını kullanarak sınırlı sayıda kelime ile bilgi arayan kullanıcıların yanında, belirli konulara odaklanmış ve bu özel konularla ilgili olarak, Internet üzerinde mevcut ya da gelecekte olması muhtemel bilgiler ile ilgilenen kullanıcıların sayısı da azımsanamayacak kadar çoktur (Liu ve Chang, 2013). Bu kapsamda geleneksel bilgi erişim sistemleri ve arama motorlarının çözüm üretemediği, bilgi erişim problemleri konusundaki çalışmalar son yıllarda ağırlıklı olarak "Konu Algılama ve İzleme (Topic Detection and Tracking-TDT)" programı içerisinde yoğunlaşmıştır. Bununla birlikte, geleneksel bilgi erişim sistemleri ve arama motorlarında karşılaşılan bilgi erişim problemleri, TDT için de popüler araştırma konuları olarak karşımızda durmaktadır.

Bu çalışmanın ana konusunu, TDT programı içerisinde tanımlanmış olan hikâye bağlantı algılama (story link detection) ve konu izleme (topic tracking) görevleri oluşturmaktadır. Bu kapsamda, bu çalışmanın temel amacı; Türkçe bir derlem üzerinde hikâye bağlantı algılama ve konu izleme görevleri için, erişim fonksiyonu ve belge gösterimi tarafında farklı yöntemler kullanarak, erişim başarımının artırılmasını sağlamaktır. Bu amacın gerçekleştirilmesi ile Türkçe etkin bir konu izleme sistem mimarisi için temel bileşenlerin de belirlenebileceği öngörülmektedir.

Belirlenen bu amacı gerçekleştirmek için, hikâye bağlantı algılama görevi ile ilgili olarak belge gösterimi bacağında belgeleri ifade etmek için seçilmesi gereken en uygun terim sayıları, erişim fonksiyonu tarafında ise, vektör uzayı ve ilgi modeli ile bunların mantıksal birleşimlerinden elde edilen sonuçlarla, en başarılı yöntemin belirlenmesi hedeflenmektedir. Buna ek olarak, konu izleme görevi ile ilgili ağırlıklı olarak erişim fonksiyonu tarafında, uygun eşik değer belirleme yöntemlerinin belirlenmesi ile vektör uzayı, ilgi modeli ve k-ortalamalar algoritmalarından başarımı en yüksek olanın belirlenmesi hedeflenmektedir.

Bu araştırma kapsamında gerçekleştirilecek olan deneysel çalışmalar sayesinde, TDT bağlamında, özellikle çok az sayıda çalışmanın bulunduğu Türkçe bir derlem üzerinde, erişim başarımının artırılması ve bu konuda bundan sonra yapılacak çalışmalara ışık tutulması amaçlanmaktadır. Bunun yanında, hikâye bağlantı algılama ve konu izleme görevi ile ilgili olarak gerçekleştirilecek başarım testlerinden elde edilen sonuçlar ışığında, eğitim belgelerinin bulunmadığı durumlarda Türkçe için, etkin bir konu izleme sistemi mimarisinin önerilmesi de bu çalışmanın amaçları içerisinde yer almaktadır.

1.2. KAPSAM

Bu çalışmanın temel kapsamını, TDT programı ile ilgili çalışmalar oluşturmaktadır. TDT programı, geleneksel bilgi erişim sistemlerinin bir devamı olarak, ilk kez 1997 yılında Amerika İleri Savunma Araştırma Projeleri (US Government's Defense Advanced Research Projects Agency, DARPA) ve Ulusal Standartlar ve Teknoloji Enstitüsü (National Institute of Standarts and Technology, NIST) tarafından başlatılan

bir çalışmanın parçası olmuş ve düzenli olarak her yıl tekrarlanan değerlendirme toplantıları ile bu konudaki aktif çalışmalar ve gelişmeler izlenmeye başlanmıştır.

TDT çalışmalarının amacı; haber yayınlarının izlenerek, sisteme içeriği yeni bir haber ulaştığında, ilgililerin uyarılmasını sağlayacak sistem ve teknolojilerin geliştirilmesini sağlamaktır. Belirlenen bu amacı gerçekleştirmek için, TDT çalışmaları, sisteme ulaşan haber yayınlarını, her biri bağımsız bir olayı tartışacak şekilde ayırmayı amaçlayan "*Hikâye Bölümleme"*, sisteme ulaşan haberin daha önce karşılaşılmamış yeni bir haber olduğunu belirlemeyi amaçlayan "*İlk Hikâye Algılama"*, sisteme ulaşan haberin hangi konu kümesine ait olduğunu belirlemeyi amaçlayan "*Küme Algılama"*, belirlenen bir haberin sistem tarafından takip edilmesini amaçlayan "*Konu İzleme"* ve sisteme ulaşan iki bağımsız haberin aynı konuyu tartışıp tartışmadıklarını anlamayı amaçlayan "*Hikâye Bağlantı Algılama"* isimleri altında beş temel göreve bölünmüştür.

Bu bağlamda, TDT görevlerinden hikâye bağlantı algılama ve konu izleme görevlerinin, Türkçe bir derlem üzerinde başarımlarının test edilmesi, bu çalışmanın temel kapsamını oluşturmaktadır. Gerçekleştirilen başarım testleri, BilCol-2005 derleminde bulunan haberlerle sınırlıdır. Testler, hikâye bağlantı algılama görevi için derlemde bulunan tüm haberler (209.305 adet), konu izleme görevinde ise, 80 konu başlığı altında sınıflandırılmış olan 5.882 haber kullanılarak gerçekleştirilmiştir.

1.3. ÖZGÜN DEĞER

Geleneksel bilgi erişim sistemleri üzerindeki akademik çalışmalar 1997 yılından sonra ağırlıklı olarak TDT programı üzerinde yoğunlaşmış ve özellikle 2000 yılında gerçekleştirilen toplantılardan sonra hikâye bağlantı algılama ve konu izleme görevlerine, bu görevlerin kritik yapısından dolayı, özel bir önem verilmiştir (Allan, 2002).

Geleneksel bilgi erişim sistemlerinden farklı olarak, TDT programında, kullanıcı sorgularının yerini, derlemdeki belgelerle ilgili olup olmadığı bilinmeyen yeni belgeler almaktadır. Bu kapsamda, hem hikâye bağlantı algılama hem de konu izleme görevlerinin gerçekleştirilmesinde, sorgu-belge eşleşmelerinin yerini, belge-belge

eşleşmeleri almakta ve bu eşleşmelerin tespiti için, geleneksel bilgi erişim sistemlerinde kullanılan yöntemler yaygın olarak kullanılmaktadır (Allan, 2002).

TDT alanında gerçekleştirilen akademik çalışmalar, özellikle erişim fonksiyonu bacağında, farklı yöntemler birlikte kullanılarak, belge gösterimi tarafında ise belgeyi ifade etmek için seçilecek uygun terimler ve terim sayıları bulunarak erişim başarımının artırılması konularında yoğunlaşmıştır (Can ve diğerleri, 2010; Yang ve diğerleri, 2002; Hatzivassiloglou, Gravano ve Maganti, 2000; Kumaran ve Allan, 2004; Kumaran ve Allan, 2005). TDT alanında da, çoğu araştırmacı, arama için seçilen kelimeler, bu kelimelerin ağırlıklandırılması ve ağırlıklandırılmış olan kelimelerin en etkili biçimde karşılaştırılması konularına odaklanmışlardır. Ancak, erişim başarımını artırmak için uygulanan her bir yöntemin, başarım üzerinde olumlu etkilerinin yanında, olumsuz etkileri de olmakta, bu nedenle, konuyla ilgili çalışmalar günümüzde halen popülerliğini korumaktadır.

TDT alanında, hikâye bağlantı algılama görevinin gerçekleştirilmesinde, farklı belge gösterim yöntemlerinin ve farklı erişim fonksiyonlarının kullanılması ve elde edilen sonuçların farklı birleşimlerinin test edilmesi konusu, literatürde yoğun olarak çalışılan bir konudur. Bunun yanında, konu izleme ile iglili olarak, Türkçe derlemler üzerinde kapsam olarak benzer çalışmalar, ağırlıklı olarak metin filtreleme, kümeleme ve sınıflandırma konularına yoğunlaşmıştır (İlhan, 2001; Kurt, 2001; Vural, 2002; Can, Altıngövde ve Demir, 2004). Tüm bunların yanında, TDT konusunda, Türkçe derlemler üzerinde gerçekleştirilen çalışmalar oldukça sınırlıdır (Can ve diğerleri, 2010; Bağlıoğlu, 2009; Can ve diğerleri, 2008; Kardaş, 2009; Acun, Başpınar, Oğuz, Saraç ve Can, 2013; Aksoy, Can ve Kocberber, 2012).

Bu çalışma kapsamında, başarım testlerinde uygulanan vektör uzayı modeli, ilgi modeli ve k-ortalamalar algoritması, geçmişten günümüze genelde bilgi erişim sistemlerinde, özelde TDT araştırmalarında, erişim fonksiyonu olarak yoğun bir şekilde kullanılmıştır (Lavrenko ve diğerleri, 2002; Allan, Carbonell, Doddington, Yamron ve Yang, 1998; Allan, 2002; Leek, Schwartz ve Sista, 2002). Bu çalışmalarda ilgi modeli kullanılarak, hem dil modeli hem de vektör uzayı modelinden daha başarılı sonuçlar alındığı gösterilmesine rağmen, farklılığı yaratan etkenler üzerinde herhangi bir yorum

bulunmamaktadır (Lavrenko ve diğerleri, 2002). Benzer pek çok çalışma, bu alanda uygulanan bir yöntemi bir diğerine göre daha başarılı olarak gösterirken, yöntemler arasındaki başarım farkının nerelerden kaynaklandığı konusunda, ayrıntılı bir çalışma gerçekleştirilmemiştir.

TDT konusunda, Türkçe derlemler üzerinde gerçekleştirilen çalışmaların son derece sınırlı olması, özellikle hikâye bağlantı algılama görevinin başarımı ile ilgili olarak hiç çalışma bulunmaması, konu izleme ile ilgili olarak ise sınırlı sayıda çalışma olması, bu çalışmayı özgün kılmaktadır. Bunun yanında, Türkçe bir derlem üzerinde geleneksel bilgi erişim sistemlerindeki sorgu-belge eşleştirme senaryolarının dışında belge-belge eşleştirmelerinde farklı yöntemlerin başarımları, belgeleri ifade etmek için seçilmesi gereken uygun terim sayıları, uygun eşik değer belirleme yöntemleri, farklı yöntemlerin mantıksal birleşimleri ile elde edilen anma (recall) ve duyarlık (precision) değerlerinin yorumlanması ve bu yorumlara göre etkin bir konu izleme mimarisinin önerilmesi, bu çalışmanın özgün değerini oluşturmaktadır.

Bu çalışma ile elde edilecek sonuçların, mükemmel bir bilgi erişim sistemine ulaşmak için ihtiyaç duyulan *"ilgili belgelerin tamamına erişim sağlama, ilgisizleri ise dışarda bırakma"* prensibine bizleri biraz daha yaklaştırması beklenmektedir.

1.4. ARAŞTIRMA PROBLEMİ VE HİPOTEZLER

Bu araştırmanın temel problemi; TDT programında tanımlı olan konu izleme görevinin, özellikle Türkçe sistemler üzerinde etkin bir şekilde gerçekleştirilememesidir. Bu problemin çözümüne katkı sağlayabilmek ve özellikle somut kanıtlara dayanan bir konu izleme sistemi mimarisi önerebilmek amacıyla, bu çalışma kapsamında aşağıdaki hipotezler test edilmektedir.

- Hikâye bağlantı algılama görevinde belgeleri göstermek için kullanılan terim sayısı arttıkça f-ölçü başarımı da artar.

- Hikâye bağlantı algılama görevinde erişim fonksiyonu olarak vektör uzayı modeli ve ilgi modelinin OR birleşimlerinin birlikte kullanılması, modellerin tek başlarına kullanıldığı yaklaşıma göre daha yüksek anma değeri sağlar.

- Hikâye bağlantı algılama görevinde erişim fonksiyonu olarak vektör uzayı modeli ve ilgi modelinin AND birleşimlerinin birlikte kullanılması, modellerin tek başlarına kullanıldığı yaklaşıma göre, daha yüksek duyarlık değeri sağlar.

- Konu izleme görevinde kümeleme için, eşik değeri olarak eğitim kümesinde "anma ve duyarlığın en yüksek olduğu değerin seçildiği yöntemin" kullanılması, "küme merkezi vektörüne eğitim belgelerinin uzaklığını temel alan yöntemlere" göre daha yüksek f-ölçü başarımı elde edilmesini sağlar.

- Konu izleme görevinde erişim fonksiyonu olarak, kümeleme tabanlı bir yöntemin kullanılması, vektör uzayı ya da ilgi modelinin kullanıldığı yönteme göre, daha yüksek f-ölçü başarımı elde edilmesini sağlar.

1.5. ÇALIŞMANIN BÖLÜMLERİ

Bu çalışma temel olarak beş bölümden oluşmaktadır.

Birinci bölümde; bu çalışmanın amacı, kapsamı, özgün değeri, araştırma problemi ve hipotezler alt başlıkları altında bilgiler sunulmaktadır.

İkinci bölümde; bu çalışmanın alanına giren ilgili çalışmalar incelenmiş olup, bilgi erişim sistemleri, konu algılama ve izleme sistemleri, hikâye bağlantı algılama ve konu izleme başlıklarında literatürdeki ilgili çalışmalara vurgu yapılmıştır.

Üçüncü bölümde; hikâye bağlantı algılama ve konu izleme görevleri için başarım testleri gerçekleştirilirken kullanılan vektör uzayı, ilgi modeli ve k-ortalamalar yöntemlerinden bahsedilmiştir. Bu bölümde ayrıca, test derlemi, test senaryoları, kullanılan araçlar ve sonuçların değerlendirilme yöntemleri, ayrıntılı olarak anlatılmıştır.

Dördüncü bölümde; hikâye bağlantı algılama ve konu izleme görevleri testerinden elde edilen sonuçlar sunulmuş, elde edilen somut sonuçların değerlendirmelerine yer verilmiş ve bu sonuçlar ışığında tez çalışması kapsamında ortaya konulan konu izleme mimarisi verilmiştir.

Beşinci ve son bölümde ise; elde edilen bulgular yorumlanmış, sonuçlar üzerindeki tartışmalar verilmiş, önerilen konu izleme sistemi mimarisi değerlendirilmiş ve gelecekte yapılması gereken çalışmalar belirtilmiştir.

2. BÖLÜM

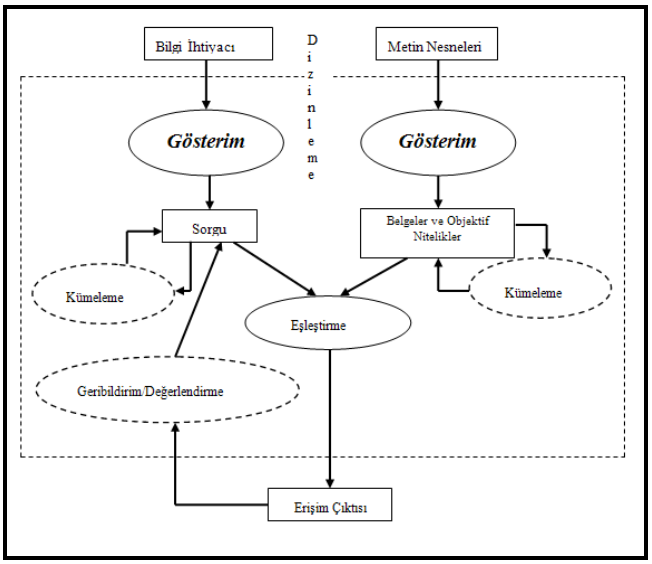
İLGİLİ ÇALIŞMALAR

2.1. BİLGİ ERİŞİM SİSTEMLERİ

Bilgi teknolojilerindeki hızlı ilerlemelere paralel olarak, kullanıcıların Internet'e olan ilgileri de ciddi oranda artmıştır. Araştırmalara göre; kullanıcıların Internet'i tercih etmelerindeki en önemli neden, bilgi ihtiyaçlarını karşılamaktır (Gordon ve Pathak, 1999). Bu kapsamda, mimari yapılarını geleneksel bilgi erişim sistemleri üzerine kuran arama motorları, kullanıcılar için temel başvuru kaynağı niteliğine bürünmüştür (Gaines ve diğerleri, 1997; Lawrence ve Giles, 1998). Ancak, Internet üzerindeki bilginin devasa boyutlara ulaşmasıyla birlikte, kullanıcıların bilgi gereksinimlerini karşılaması beklenen arama motorları yetersiz kalmaya başlamıştır (Voorbij, 1999; Balabantaray ve diğerleri, 2013). Kullanıcıların Internet üzerinde ilgili bilgiye erişim konusunda yaşadıkları problemler, geleneksel bilgi erişim sistemlerinin mantıksal organizasyonundan (Maron, 1984) kaynaklanmaktadır.

Bilgi erişim sistemleri, farklı ortamlarda bulunan belgeler içerisindeki bilginin bulunarak, onunla ilgilenen kullanıcılara sunulmasını amaçlayan sistemlerdir (Meadow, 1992). Bir bilgi erişim sistemi, belgelerin bulunduğu derlem, kullanıcı sorguları ve kullanıcıların sorgu cümlelerinde yer alan terimlerle, derlemdeki belgelere verilen terimleri karşılaştırarak, ilgili belgeleri belirlemek için kullanılan bir erişim fonksiyonundan oluşur. Bu noktada, bilgi erişim sisteminin temel işlevi, kullanıcıların bilgi ihtiyaçlarını karşılaması muhtemel derlemdeki ilgili (relevant) belgelerin tümüne erişmek, ilgili olmayanları da ayıklamaktır (Tonta, Bitirim ve Sever, 2002).

Tonta ve diğerleri (2002); geleneksel bilgi erişim sistemlerinin işlevsel mimarisini, Maron'un (1984) tanımını bir adım daha ileriye götürerek, üçer adet ön ve arka yüz kavramı ile açıklamışlardır. Araştırmacılar, ön yüz kavramlarını sistemin dış dünyaya yansıyan görünüşü, arka yüz kavramlarını da bilgi erişim süreçleri arasındaki iletişimde kullanılan bileşenler olarak tanımlamışlardır.



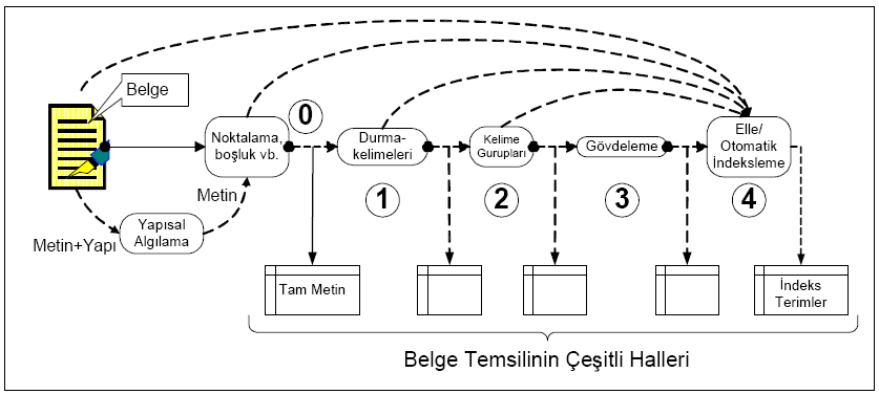
Şekil 1. Bir bilgi erişim sisteminin işlevsel mimarisi (Kaynak: Tonta, Bitirim ve Sever 2002:12)

Şekil 1'de gösterilen bu işlevsel mimari içerisinde bilgi ihtiyacı, metin nesneleri ve erişim çıktısı önyüz; sorgular, belgeler ve içerik belirteçleri de arka yüz kavramlarını oluşturmaktadır. Sistemin işleyişi kısaca şu şekilde gerçekleşmektedir: Bir yanda, doğal dille ifade edilen kullanıcı bilgi ihtiyacı, sistem tarafından sorgu terimlerine dönüştürülerek eşleştirme fonksiyonunun ilk girdisini oluşturur. Diğer tarafta ise, sistem tarafından sürekli olarak dizinlenerek ters dizin kütüklerinde (inverted index) tutulan belgeler, eşleştirme fonksiyonunun diğer girdisini oluşturmaktadır. Eşleştirme fonksiyonu kullanıcı sorgusu ile derlemde bulunan belgeleri karşılaştırarak, derlemde sorgu ile ilgili belgeleri belirler. Belirlenen belgeler, sorgu ile ilgililik derecelerine göre, en ilgili belgeden başlanarak sıralanır ve bu ilgili belgelerin sıralı listesi, erişim çıktısını oluşturur. Sistemin işleyişi esnasında erişim fonksiyonunun sorgularla belgeleri eşleştirme sürecini hızlandırmak için belgeler, bilgi erişim performans etkinliğini artırmak için de sorgular kümelenebilir (Tonta ve diğerleri, 2002; Lee, 1995; Belkin, Kantor, Fox ve Shaw, 1995). Diğer taraftan, erişim çıktısının, kullanıcının bilgi ihtiyacını karşılamaktan uzak olması durumunda, daha ilgili çıktılara erişebilmek amacıyla, kullanıcı tarafından bir ilgililik geribildirim (relevance feedback) sürecinin başlatılabilmesi de mümkündür (Salton ve Buckley, 1990).

Bilgi Erişim Sistemlerinin mantıksal organizasyonu içerisinde, bir tarafta sistem tarafından derlem içerisinden çekilerek dizinlenen belgeler bulunurken, diğer tarafta

bilgi ihtiyacını karşılamayı amaçlayan kullanıcılar ve bu kullanıcıların bilgi ihtiyaçlarını ifade ettikleri sorgu cümleleri bulunmaktadır. Bu sistem içerisindeki en kritik bileşen, sorgu cümleleri ile dizin terimleri arasındaki çakışmalara göre, sorgu ve belgeler arasındaki benzerlik değerlerini belirleyen erişim fonksiyonudur (Tonta, 1995). Erişim fonksiyonu, doğası gereği, yalnızca kullanıcıların sorgu cümlelerinde geçen ve dizin terimleri ile kesişen belgelere erişim sağlar.

Bilgiye erişimde, erişim fonksiyonunun etkinliği kadar bu aşamadan önce uygulanan ön işlemlerin de etkisi büyüktür. Dinçer (2004) bu süreci; bilgi erişim sistemlerinde yaygın olarak kullanılan biçimi ile Şekil 2'de gösterildiği gibi ifade etmiştir. Şekilde; kesikli oklar seçimlik, kesiksiz oklarsa zorunlu alt süreç işlemlerine akışını yönlendirmektedir.



Şekil 2. Bilgi erişim sistemlerinde belge gösterimi süreci (Kaynak: Dinçer 2004:87)

Belge gösterim sürecini oluşturan alt süreçler, doğal dille ifade edilmiş bir belgeden başlayarak, sonunda dizin terimlerine ulaşılması ile son bulan bir dönüşüm zinciri gibi düşünülebilir. Doğal dille ifade edilen bir belgenin üzerinde gerçekleştirilecek asgari düzeydeki belge gösterimi, ilgili belgeden boşluklar, noktalama işaretleri gibi anlamsal olarak herhangi bir değeri olmayan simgelerin çıkarılarak belgenin, şekilde gösterildiği gibi, *"tam metin (full text)"* olarak ifade edilmesi ile gerçekleştirilir. Geleneksel bilgi erişim sistemlerinde, bu işlem metin normalleştirme (text normalization) ya da simgeleştirme (tokenization) olarak adlandırılmaktadır. Bu işlem aynı zamanda, belge gösterim sürecinde uygulanan sonraki yordamlar için de bir girdi olarak kabul edilmektedir.

Belge gösterim sürecindeki sonraki yordam ise durma kelimelerinin (stop words) çıkarılması işlemi ile ifade edilmektedir. Durma kelimeleri, bir metin içerisinde asıl ifade edilmek istenen konuyla çok ilgili olmayan ve metni diğer metinlerden ayırdedici özelliği bulunmayan kelimeler (ve, veya, ile, ise vb.) olarak kabul edilmektedir. Durma kelimeleri, doğal dil içerisinde 200-300 kelimeyi geçmemesine rağmen, bu kelimelerin metin içerisindeki kullanım sıklıkları oldukça yüksektir.

Belge gösterim sürecinin bir sonraki yordamını oluşturan gövdeleme işlemi, uzun süredir araştırmacıların ilgisini çekmekte ve BES alanında önemli olarak görülmektedir (Frakes ve Yates, 1992). Gövdeleme; kelimeden çekim eklerinin çıkarılarak yapım eklerinin bırakılması olarak tanımlanabilir. Analitik diller üzerinde yapılan bazı çalışmalarda, gövdelemenin bilgi erişim performansını olumlu yönde etkilediği savunulurken, pek çok çalışmada bunun tam tersi sonuçlar elde edilmiştir (Krovetz, 1993; Hull, 1996; Sheridan ve Balerini, 1996; Popovic ve Willet, 1992). Örneğin, Harman (1991) İngilizce belgeler üzerinde, farklı gövdeleme algoritmalarını uygulayarak erişim etkinliğini artırmaya çalışmış ancak başarılı olamamıştır. Benzer şekilde, İspanyolca için yapılan bir çalışmada, gövdeleme işleminin erişim etkinliğini artırmaya yönelik bir etkisi tespit edilememiştir (Figuerola, Gomez, Rodriguez ve Berrocal, 2002). Diğer taraftan, bazı araştırmalarda, Almanca ve İngilizce için gövdelemenin erişim etkinliğini artırdığı rapor edilmiştir (Braschler ve Ripplinger, 2004; Hull, 1996; Krovetz, 1993).

Gövdeleme, özellikle Türkçe belgeler için geliştirilen bilgi erişim sistemlerinde de önemli araştırma sorularından birisi olmuştur. Bu konuda, Türkçe derlemler için ilk çalışma Köksal (1981) tarafından gerçekleştirilmiş ve gövdeleme işlemi kelimelerin ilk 5 harfleri alınarak yapılmıştır. Sonraki yıllarda, Solak ve Can (1994) 533 haber ve 71 sorgu kullanarak, Türkçe gövdelemenin erişim etkinliğine etkisini araştırmışlar ve %9'a kadar etkinlik artışından bahsetmişlerdir. Ekmekçioğlu ve Willett (2000) ise 6289 belge ve 50 sorgu kullanarak, yeni bir test gerçekleştirmişler ve erişim başarımının erişim çıktısındaki ilk 10 ve 20 belge için %32 oranında arttığını rapor etmişlerdir. Sonraki yıllarda Sever ve Bitirim (2003) daha önce hiç kullanılmamış olan "*gövde bul*" isimli bir algoritma geliştirerek 2468 hukuk belgesi içinde 15 sorgu ile bu algoritmayı test

etmişlerdir. Bu çalışmada, Türkçe için o güne kadar rapor edilen en iyi sonuçlara ulaşılmış ve %25'lik bir başarım artışı rapor edilmiştir.

Diğer taraftan, son yıllarda Türkçe belgeler üzerinde gerçekleştirilen çalışmalarda, önceki çalışmaların tersine (Solak ve Can, 1994; Ekmekçioğlu ve Willett, 2000; Sever ve Bitirim, 2003) gövdelemenin bilgi erişimin başarımında anlamlı bir artış etkisi göstermediği savunulmaktadır (Tunalı ve Bilgin, 2012; Torunoglu, Cakirman, Ganiz, Akyokus ve Gurbuz, 2011). Tunalı ve Bilgin (2012), çeşitli Türkçe haber sitelerinden toplanan haberlerin, otomatik olarak kümelenmesi için yaptıkları çalışmada, farklı gövdeleme algoritmalarını deneyerek sonuçları değerlendirmişlerdir. Araştırmacılara göre, kümeleme işleminde Türkçe gövdeleme, anlamlı bir başarım artışı sağlamamakla birlikte, oluşturulan dizin boyutunun ciddi oranda küçültülmesine yardımcı olmuştur (Tunalı ve Bilgin, 2012). Torunoglu ve diğerleri (2011) tarafından gerçekleştirilen çalışmada, web üzerinden Türkçe haber sitelerinden toplanan haberlerin otomatik sınıflandırılmasında, metinler üzerinde gerçekleştirilen ön işlemenin başarım üzerindeki etkisi araştırılmış ve hemen hemen hiç etkisi olmadığı rapor edilmiştir.

Geleneksel bilgi erişim sistemlerindeki etkinlik probleminin çözümüne yönelik olarak gerçekleştirilen çalışmalar, kullanıcı tarafında sorgu cümlelerinin genişletilerek, sorgu terimlerinin dizin terimleri ile çakışma olasılıklarının, erişim fonksiyonu tarafında da farklı yöntemler kullanılarak erişim etkinliğinin artırılması biçiminde kendini göstermiştir. Sorgu genişletme konusunda alana özel başlıklar, kavramsal sözlükler, belge analizi yöntemleri, kavramsal ilişki tabanlı yöntemler ve alan ontolojileri tabanlı yöntemler kullanılmaktadır (Chen ve diğerleri, 1995; Xu ve Croft, 1996; Song, Song, Hu ve Allen, 2007; Bhogal, Macfarlane ve Smith, 2007). Erişim fonksiyonu konusunda ise kullanılan yöntemler genel olarak; Boole modeli (Robertson, 1977), vektör uzayı modeli (Salton, Wong ve Yang, 1975), olasılıksal modeller (Robertson, 1977; Maron, 1988; Maron ve Kuhns, 1960; Sparck Jones, Walker ve Robertson, 2000), dil modeli (Ponte ve Croft, 1998) ve ilgi modeli (Lavrenko ve Croft, 2001) olarak karşımıza çıkmaktadır.

2.2. KONU ALGILAMA VE İZLEME

Günümüzde yeni teknolojilerin gelişmesi ile birlikte, Internet kullanıcılarının bu mecradan beklentileri de artmıştır. Elektronik ticaretin yoğunlaşması, sosyal medyanın popülaritesinin artması ve Internet üzerinden yayın yapan haber kaynaklarının çoğalması ve çeşitlenmesi bir şekilde bu beklentilerin ürettiği sonuçlar olarak görülebilir.

Internet üzerinde sayıları ve çeşitleri zaman içerisinde hızla artan haber siteleri, güncel haberleri kullanıcılara gerçek zamanlı olarak sunarken, bu yetenek dünyada olup bitenden haberdar olmak isteyen kullanıcıların, akşam haberlerini ya da ertesi gün çıkacak gazeteleri beklemelerini gereksiz hale getirmiştir. İşte bu avantaj, Internet üzerinden yayın yapan haber sitelerini, bu mecranın en popüler bilgi varlıklarından birisi haline dönüştürmüştür (Liu ve Chang, 2013). Buna karşılık bu kadar çok haber sitesini takip etmek olanaksız hale gelmiş ve haberlerin gün içindeki hızlı akışında ilgi duyulan pek çok bilgi gözden kaçmaya başlamıştır.

Internet kullanıcılarının gündemdeki hızlı akışı hiçbir ayrıntıyı kaçırmadan takip edebilme gereksinimleri, geleneksel bilgi erişim sistemleri ve arama motorlarının işlevlerini gözden geçirme ihtiyacı oluşturmuştur. Bu kapsamda geleneksel bilgi erişim sistemlerinin mimari yapısı üzerine kurulan arama motorlarını kullanarak, sınırlı sayıda kelime ile bilgi arayan kullanıcıların yanında, belirli konulara odaklanmış ve bu özel konularla ilgili olarak, Internet üzerinde mevcut ya da gelecekte olması muhtemel bilgiler ya da haberler ile ilgilenen kullanıcıların sayısı da azımsanamayacak kadar artmıştır (Liu ve Chang, 2013).

Kullanıcıların bu tür bilgi ihtiyaçlarına çözümler üretmek için başlamış olan TDT programında, özellikle hikâye bağlantı algılama ve konu izleme görevleri kritik bileşenler olarak kabul edilmiştir (Allan ve diğerleri, 1998). Çalışmanın bu bölümünde, bu iki görevle ilgili olarak gerçekleştirilmiş olan araştırmalara vurgu yapılarak, mevcut durum ortaya konulmaktadır.

2.2.1. Hikâye Bağlantı Algılama

TDT çalışmaları, her ne kadar geleneksel bilgi erişim sistemlerinin bir devamı olarak düşünülse de işlevsel mimari açısından bakıldığında, bazı farklılıklar göze çarpmaktadır. Geleneksel bilgi erişim sistemlerinde, bilgi ihtiyacı, genellikle kullanıcı tarafından ifade edilen kelimelerden ya da sınırlı sayıda cümleden oluşmaktadır. Tam bu noktada, geleneksel bilgi erişim sistemleri ile TDT çalışmaları birbirinden ayrılmaktadır. TDT içerisinde bilgi ihtiyacı, genellikle belirli bir konuyu tartışan ve klasik kullanıcı sorgularından çok daha fazla sayıda içerik terimi barındıran belgeler olarak düşünülür. TDT içerisinde ilgililik değerlendirmesi yapılırken, kullanıcı sorgusu yerine aynı konuda olup olmadığı merak edilen belgeler birbiri ile karşılaştırılır.

TDT programı içinde, hikâye bağlantı algılama görevi, sisteme veri kaynaklarından ulaşan hikâyeler içinde, hangi hikâye çiftlerinin aynı konuyu tartıştıklarını tespit etmeye çalışır ve her bir hikâye çifti için "evet" veya "hayır" yanıtları üretir. Bu kapsamda, sistem tarafından verilen kararların, ne derece güvenilir olduğunu belirlemek amacıyla haber çiftleri için bir skor değeri üretilir. Daha sonra, bu skor değerleri içerisinden bir eşik değeri seçilerek, bu değerin üzerindeki skor değerlerine sahip haber çiftleri aynı konu üzerinde, eşik değerinden düşük skora sahip haber çiftleri de farklı konularda olarak kabul edilir (Martin, Doddington, Kamm, Ordowski ve Przybocki, 1997; Fiscus, Doddington, Garofolo ve Martin, 1999).

Bu kapsamda, hikâye bağlantı algılama görevinin, TDT çalışmalarında kritik bir öneme sahip olduğu belirlenmiştir (Lavrenko ve diğerleri, 2002; Allan ve diğerleri, 1998; Allan, 2002). Sisteme verilen iki bağımsız haberin, aynı konuyu tartışıp tartışmadığını anlamayı hedefleyen hikâye bağlantı algılama görevinin başarıyla gerçekleştirilmesi halinde, TDT için pek çok problemin de beraberinde çözülebileceği öngörülmektedir (Allan ve diğerleri, 1998; Allan, 2002).

Hikâye bağlantı algılama görevinde, geleneksel bilgi erişim sistemlerinden farklı olarak, iki farklı belgenin aynı konuda olup olmadığı belirlenmeye çalışılmaktadır. Bu kapsamda, bilgi erişim sistemlerinde de kullanılan boole modeli, vektör uzayı modeli,

olasılıksal modeller, dil modeli ve ilgi modeli gibi pek çok erişim fonksiyonu yönteminin TDT içerisinde de kullanıldığı görülmektedir.

Bunun yanında, TDT kapsamında gerçekleştirilen çalışmalar, ağırlıklı olarak erişim başarımının artırılmasına yönelik olarak farklı yöntemlerin uygulanması konusunda yoğunlaşmıştır. Bu kapsamda, TDT programı içerisinde sistem başarımını artırabilmek için, belge gösterimi ve farklı erişim fonksiyonlarının sonuçlarının birleştirilmesi ile ilgili yöntemler yoğun olarak araştırılmıştır (Salton, 1989; Ponte ve Croft, 1998; Thompson ve Callan, 2005; Shah, Croft ve Jensen, 2006; Kumaran ve Allan, 2004; Kumaran ve Allan, 2005; Can ve diğerleri, 2010; Allan, Lavrenko ve Jin, 2000; Makkonen, Ahonen ve Salmenkivi, 2003; Makkonen, Ahonen ve Salmenkivi, 2002; Qiu, Liao ve Dong, 2008; Qiu ve Liao, 2008; Mori, Miura ve Shioya, 2006; Jin, Myaeng, Lee, Oh ve Jang, 2005; Kim ve Myaeng, 2004; Can ve diğerleri, 2010; Yang ve diğerleri, 2002; Hatzivassiloglou ve diğerleri, 2000; Kumaran ve Allan, 2004; Kumaran ve Allan, 2005; Köse, Tonta, Ahmadlouei ve Polatkan, 2013; Soydal ve Al, 2014).

Erişim fonksiyonu tarafında, farklı yöntemlerin birleştirilmesi konusunda yapılan çalışmalar (Can ve diğerleri, 2010; Yang ve diğerleri, 2002; Hatzivassiloglou ve diğerleri, 2000; Kumaran ve Allan, 2004; Kumaran ve Allan, 2005; Köse, Tonta, Ahmadlouei ve Polatkan, 2013; Köse ve Ahmadlouei, 2013) genellikle sistemin anma değerlerini artırırken, aynı zamanda ilgisiz pek çok belgenin de getirilmesini sağlamakta ve sistemin duyarlık değerinin dolayısıyla başarımın düşmesine neden olmaktadır. Bu nedenle, bu tür farklı erişim fonksiyonlarının birlikte kullanılacağı çalışmalarda sistem başarımını en üst seviyeye çıkarabilmek için, anma ve duyarlık arasındaki dengeyi gözetecek modellerin geliştirilmesi son derece önemlidir. Kısacası bu tür sistemlerin ideal olarak derlemdeki tüm ilgili belgelere erişim sağlamasını, aynı zamanda da ilgisizlerin dışarıda bırakılmasını sağlayacak şekilde uygun stratejileri desteklemesi gerekmektedir.

Hikâye bağlantı algılama görevinin gerçekleştirilmesinde kullanılan pek çok yöntem, karşılaştırılan iki hikâye arasında ne kadar fazla sayıda kelimenin örtüştüğünü araştırır. Karşılaştırılan iki hikâye arasında ne kadar fazla sayıda örtüşen kelime varsa, bu iki

hikâyenin aynı konuyu tartışma olasılığının da o kadar yüksek olduğu kabul edilir. Bu yaklaşım, vektör uzayı modellerinden (Frakes ve Baeza,1992; Allan, Lavrenko ve Swan, 2002; Schultz ve Liberman, 1999; Schultz ve Liberman, 2002; Xu ve Croft, 2000; Ponte ve Croft, 1997) başlayıp, istatistiksel dil modellerine kadar (Berger ve Lafferty, 1999; Miller, Leek ve Schwartz, 1999; Song ve Croft, 1999; Ponte ve Croft, 1998; Lavrenko ve Croft, 2001) geliştirilen bütün yöntemlerin temelini oluşturmuştur.

Chen, Farahat ve Brants (2004) yaptıkları çalışmada; hikâye bağlantı algılama görevinin gerçekleştirilmesi için farklı erişim fonksiyonları ve bunların birleşimlerini kullanarak TDT-2002 derlemi üzerinde başarım testleri gerçekleştirmişlerdir. Belge gösterimlerinde *tf.idf* yaklaşımını temel alan araştırmacılar, haberleri karşılaştırırken kosinüs, normalize kosinüs, Hellinger, Tanimoto ve clarity benzerlik ölçüm yöntemlerini ve bunların farklı birleşimlerini kullanarak testler gerçekleştirmişlerdir. Araştırmacılar bu çalışmada istatistiksel yöntemlerle (clarity) desteklenen temel yöntemlerde (normalize kosinüs), en iyi performansın elde edildiğini rapor etmişlerdir (Chen ve diğerleri, 2004).

Lakshmi ve Mukherjee (2007) ise; hikâye benzerliklerini belirlemek için, çalışmalarında birleşme modeli (cohesion model) adını verdikleri bir yöntem kullanmışlardır. İlgi modelinden uyarlanan bu yöntemde, her bir haber derlemde bulunan diğer ilgili haberler kullanılarak yeniden modellenmekte ve haberleri ifade etmek için kullanılan terimler, özel bir ağırlıklandırma yöntemi kullanılarak seçilmektedir. TDT-4 derlemi kullanılarak gerçekleştirilen testler sonucunda araştırmacılar, kosinüs benzerliği yöntemine göre daha yüksek bir başarım elde edildiğini belirtmişlerdir (Lakshmi ve Mukherjee, 2007).

Bir diğer çalışmada ise Nomoto (2010); haber benzerliklerini belirlemek için belge benzerlikleri ve kullanıcı geri beslemelerinden oluşan iki katmanlı bir model geliştirmiştir. İlk katmanda, haberler farklı erişim fonksiyonları kullanılarak karşılaştırılmış ve birinci benzerlik skorları elde edilmiştir. İkinci katmanda ise, haberler birer sorgu olarak kullanılmış, derlemde bulunan diğer ilgili belgeler kullanılarak konu modelleri yaratılmış ve yaratılan konu modelleri "clarity yöntemi" kullanılarak karşılaştırılmıştır. Bu karşılaştırma sonucu elde edilen skorlar, ilk katmanda elde edilen skorlarla birleştirilerek TDT-5 derlemi üzerinde başarım testleri gerçekleştirilmiştir.

İlgili yöntem, en başarılı olarak bilinen Kullback Liebler ve clarity yöntemlerinden daha yüksek bir başarım sağlamıştır (Nomoto, 2010).

Shah ve Eguchi (2009); haber benzerliklerinin belirlenmesinde terimlerin seçilmesi konusunda, klasik bilgi erişim yöntemlerinin yeterli olmayacağını savunmuşlar, bir haberi ifade etmek için hem belgenin kendisinden hem de derlemdeki diğer belgelerden yaralanarak, en iyi terimleri belirlemeye çalışmışlardır. Çalışmada hem *tf.idf* ağırlıklandırmanın yeni bir uyarlaması hem de derlemdeki terimlerin olasılıksal dağılımlarının Kullback Leibler yönteminin farklı uyarlamaları kullanılarak haberler için konu modelleri yaratılmıştır. Araştırmacılar, elde edilen sonuçların, klasik bilgi erişim sistemlerinde kullanılan belge gösterim tekniklerinden çok daha başarılı olduğunu rapor etmişlerdir (Shah ve Eguchi, 2009).

Belge gösterimleri, hem geleneksel BES hem de TDT görevleri için son derece önemli bir aşamadır. Çalışılan alanlara bağlı olmak koşulu ile belge gösterimi için kelime tabanlı yöntemler (Salton, 1989), dil modelleri (Ponte ve Croft, 1998) ve çizge (graph) tabanlı yöntemler (Thompson ve Callan, 2005) kullanılmaktadır. Belge gösterimi ile ilgili olarak kullanılan yöntemlerden bazıları konudan bağımsız olarak geniş bir kullanım alanı bulurken, diğer bazı yöntemler sadece sınırlı alanlarda kullanılabilmiştir. TDT çalışmaları da doğası gereği, belge gösteriminin kritik bir öneme sahip olduğu alan olarak karşımıza çıkmaktadır. TDT, haber metinleri içerisinde ifade edilen olaylar (events) ile doğrudan ilgilidir ve bu program içerisinde bir olay; özel bir mekânda, belirli kişi ya da organizasyonların katılımı ile belirli bir zaman diliminde gerçekleşen eylemler olarak tarif edilmektedir (Shah ve diğerleri, 2006). Bu kapsamda, TDT içerisinde bir haber metninin gösteriminde varlık isimlerinin (named entity) kullanılması ile ilgili çalışmalar popüler araştırma konularından olmuştur.

Shah ve diğerleri (2006), çalışmalarında; TDT içerisinde tanımlı olan hikâye bağlantı algılama görevinin gerçekleştirilmesi amacıyla, haber benzerliklerinin belirlenmesinde, varlık isimlerinden yararlanmışlardır. Çalışmada *tf.idf* ağırlıklandırma yöntemi baz olarak kabul edilmiş ve bu yöntemin başarımı varlık ismi tabanlı *tf.idf*, ağırlıklandırılmamış varlık ismi genişletme yöntemi ve ağırlıklandırılmış varlık ismi genişletme yöntemleri ile karşılaştırılmıştır. Bu çalışmada varlık isimleri kullanılarak

uygulanan ilk yöntemde (*tf.idf* on entities) BBN's Identifier (Bikel, Schwartz ve Weischedel, 1999) kullanılarak varlıklar otomatik olarak tespit edilmiş ve haber metinlerinde geçen diğer kelimeler (isimlendirilmiş varlıklar dışındakiler) atılmıştır. Sonraki aşamada, her bir belge için belirlenen varlık isimleri kullanılarak belge vektörleri oluşturulmuştur. Belge benzerliklerinin belirlenmesinde, vektör uzayı modeli kullanılmıştır. Bu yöntemde, en büyük problem, bazı belgelerin sağlıklı bir karşılaştırma yapacak kadar varlık ismine sahip olmamasıdır. Bu problemi gidermek için varlıklar arasındaki ilişkileri gösteren çizgeler oluşturulmuş ve aynı haberde bir kez birlikte geçen varlık isimleri, ilişkili olarak kabul edilmiştir. Bu yaklaşımda, belge vektörleri oluşturulurken, sadece belge içinde geçen varlıklar değil, bunlarla ilişkili diğer varlıklar da kullanılmıştır (unweighted expansion). Uygulanan son yöntemde ise, çizge üzerinde birbiri ile ilişkili varlık isimlerine, ilişki derecelerine göre bazı ağırlıklar verilmiş ve yeni belge vektörleri, bu ağırlıklar göz önüne alınarak oluşturulmuştur. Testler sonucu elde edilen veriler, hikâye bağlantı algılama görevinde haber benzerlikleri belirlenirken varlık isimlerinin kullanılmasının, sistem başarımı üzerinde anlamlı bir artış sağladığını göstermiştir (Shah ve diğerleri 2006).

Varlık isimlerinin TDT programında "Yeni Olay Algılama (New Event Detection – NED)" görevi için kullanıldığı diğer bir çalışma, Kumaran ve Allan (2004) tarafından gerçekleştirilmiştir. Bu çalışmadan elde edilen sonuçlar, yeni olay algılama görevinin gerçekleştirilmesinde, varlık isimlerinin kullanılmasının, belirli konularda başarım üzerinde olumlu etkisi olduğunu göstermektedir (Kumaran ve Allan, 2004).

Bu çalışmanın devamında Can ve diğerleri (2010), Türkçe bir derlem üzerinde yeni olay algılama görevinin gerçekleştirilmesinde varlık isimlerinin sistem başarımı üzerindeki etkilerini araştırmışlardır. Araştırmada, belge vektörleri oluşturulurken dört farklı yöntem kullanılmıştır. Bu yöntemler: 1) varlık ismi dışındaki tüm kelimelerin alınması; 2) sadece varlık isimlerinin alınması; 3) tüm kelimelerin alınması ve 4) Kumaran ve arkadaşları (2004) tarafından önerilen üçgenleme (triangularization) yaklaşımıdır. Bu çalışmada, belgeler içerisindeki varlık isimlerinin belirlenmesinde, otomatik çıkarsama yöntemleri kullanılmıştır. Buna göre, belgeler içerisindeki tüm kelimelerin kullanıldığı vektör gösterimi yaklaşımı, en başarılı yöntem olarak rapor edilmiştir (Can ve diğerleri, 2010).

Geleneksel bilgi erişim sistemlerinde kullanılan belge gösterim yöntemlerinin aslında TDT için yetersiz kaldığına ve bu alanda olay tabanlı destekleyici farklı yöntemlerin kullanılması gerektiğine literatürde sıkça vurgu yapılmıştır (Allan ve diğerleri, 2000; Makkonen, Ahonen ve Salmenkivi, 2003; Makkonen ve diğerleri, 2002; Qiu ve diğerleri, 2008; Qiu ve Liao, 2008; Mori ve diğerleri, 2006; Jin ve diğerleri, 2005; Kim ve Myaeng, 2004). Bu bakış açısı ile TDT içerisindeki belgeleri, klasik terim vektörleri ile ifade etmek yerine, hikâyeler içerisindeki isimleri, yerleri, zamanı ve konuyu adresleyen olay vektörlerinin (event vectors) kullanılmasının daha anlamlı olacağı fikri destek görmüştür (Makkonen ve diğerleri, 2003). Buna göre, bir olay vektörü, olaya katılan aktörleri ifade eden kişiler (who), olayın gerçekleştiği zamanı ifade eden zaman (when), olayın gerçekleştiği mekânı ifade eden konum (where) ve olayın eylemini ifade eden konu (what) vektörlerinden oluşacak biçimde ifade edilebilir.

Kumaran ve Allan (2005), NED ile ilgili olarak gerçekleştirdikleri bir çalışmada; varlık isimlerini kullanarak iki farklı hikâyenin karşılaştırılması için isimler, konular ve tam metinleri dikkate alarak bazı deneyler yapmışlardır. Yazarlar, TDT içerisindeki olay (event) tanımından yola çıkarak, bir hikâyenin kişiler (who), yerler (where), zaman (when) ve eylemi belirleyen (what) kelimeler kullanılarak ifade edilebileceğini söylemişlerdir. Bu kabûle göre; eğer iki farklı hikâye aynı konuda ise bu hikâyelerin aynı varlık isimlerini ve konu terimlerini paylaşmaları gerekir. Diğer taraftan, eğer iki hikâye birbirine yakın ancak farklı konularda ise, varlık isimleri ya da konu terimleri arasında bir eşleşme olsa da muhtemelen her ikisi birden eşleşmeyecektir (Kumaran ve Allan, 2005). Bu çalışmada, varlık isimleri kullanılarak gerçekleştirilen sınıflandırma yöntemlerinin, vektör uzayı modeli temel alınarak gerçekleştirilen temel sınıflandırma modelinden anlamlı olarak daha başarılı sonuçlar elde edildiği rapor edilmiştir.

Benzer bir yaklaşım, daha önceleri Makkonen ve diğerlerinin (2002) çalışmalarında da kullanılmıştır. Araştırmacılar, haberlerde geçen isim, yer ve zaman bilgilerini ayrı ayrı vektörlerle ifade etmişlerdir. Bu çalışmada, isim, yer ve zaman gibi varlık isimleri otomatik çıkarsama yöntemleri ile elde edilmiş ve belge içerisinde, bunlar dışındaki terimlerin, haberin konusunu (what) ifade edeceği belirtilmiştir. Yazarlar, varlık isimlerinin kullanılmasının, yeni haber algılama probleminde önemli bir başarım artışı sağladığını rapor etmişlerdir (Makkonen ve diğerleri, 2002). Araştırmacılar takip eden

çalışmalarında (Makkonen ve diğerleri, 2003), TDT için sadece belge terimleri kullanılarak gerçekleştirilen belge gösterimlerinin yeterli olmadığını ve etkili bir sistem için, varlık isimleri kullanılması gerektiğini vurgulamışlardır. Araştırmacılar, her iki çalışmalarında da özellikle yer ve zaman karşılaştırmaları için kesişime dayanan benzerlik metrikleri önermişlerdir (Makkonen ve diğerleri, 2002, 2003).

TDT görevlerinin gerçekleştirilmesinde varlık isimlerinin kullanılmasının, literatürde genellikle başarım üzerindeki olumlu etkilerinden bahsedilmekle birlikte, bunun tersinin savunulduğu çalışmalar da vardır. Kim ve Myaeng (2004), Korece haberlerden oluşturulmuş olan derlem üzerinde gerçekleştirdikleri çalışmalarında, zaman (when) bilgisinin, konu takibi (topic tracking) için gerçekleştirilen deneylerde, başarımı anlamlı bir oranda artırmadığını ifade etmişlerdir.

2.2.2. Konu İzleme

Konu izleme sistemleri, temel olarak, kullanıcının ilgi alanında bulunan belgeler ve kullanıcı özelliklerini dikkate alarak, hedef belgelerden hangilerinin kullanıcının ilgi alanında olduğunu belirlemeye çalışır (Gupta ve Lehal, 2009). Konu izleme sistemleri, endüstride firmaların kendileri ve rakipleri ile ilgili yeni çıkan bilgileri takip etmelerinden başlayarak, tıp alanında doktorların yeni tedavi yöntemlerinden, akademik alanda bilim insanlarının çalıştıkları konulardaki son yayınlardan haberdar olmalarına kadar pek çok farklı alanda kullanılmaktadır (Kaur ve Gupta, 2012).

Internet'in hızlı gelişimi ile birlikte, belirli konularla ilgili bilgiler, farklı zaman dilimleri ve konumlara yayılmaya başlamış olup TDT çalışmaları ile, bu dağınık bilgilerin farklı konumlardan toplanarak organize edilmesi ve kullanıcılara daha rahat anlaşılır bir biçimde sunulması hedeflenmiştir (Xiaowei, Longbin ve Jialin, 2008). Bu bağlamda, haber kaynaklarından sisteme gelen yeni haberlerin değerlendirilerek, bu haberlerin daha önceden belirlenmiş olan konu ya da konularla ilgili olup olmadığını tespit etmeyi amaçlayan konu izleme sistemlerinin (Zhang, Guo ve Li, 2009) günümüz ihtiyaçları için kritik bir öneme sahip olduğu ve bu konudaki çalışmalara yoğunluk verilmesi gerektiği belirtilmiştir (Allan ve diğerleri, 2000; Allan, 2002).

Konu izleme, bilgi erişim sistemlerinde kullanılan bilgi süzme (information filtering) işlemine benzer bir yapıdadır. Sisteme aynı konuyla ilgili olduğu bilinen az sayıda (genellikle 1 ile 4 arası) haber verilerek, sistemin haber kaynağından ulaşan haberler içinden bu konuyla ilgili olan tüm haberleri yakalaması beklenir. Bu bağlamda, bir konu izleme sistemi, konu gösterimi (topic representation), haber gösterimi (story representation), benzerlik ölçümü, eşik değer karşılaştırması ve benzerlik kararı bileşenlerinden oluşur (Qin ve Zhang, 2008).

Konu izleme sistemlerinde, öncelikle izlenmek istenen konuyla ilgili olarak az sayıda haber kullanılarak, bu haberlerin öz niteliklerinden bir konu modeli oluşturulur. Benzer bir yaklaşım, karşılaştırılacak haberler için de kullanılarak her bir haber için haber öznitelikleri oluşturulur. Böylece konu ve haber gösterimleri elde edilerek sistem benzerlik ölçümü için hazır hale getirilmiş olur. Sonraki aşamada, bu gösterimleri karşılaştırmak için, bir erişim fonksiyonu kullanılarak konu ve haber benzerliği için bir skor elde edilir (Allan, 2002; Yang, Ault, Pierce ve Lattimer, 2000). Elde edilen bu skor, daha önce belirlenmiş bir eşik değer ile karşılaştırılır. Eğer elde edilen skor, belirlenmiş olan eşik değerin üzerinde ise, haber konuyla ilgili, altında ise ilgisiz olarak kabul edilir ve bu işlem, sisteme ulaşan her bir yeni haber için tekrarlanır. Konu izleme sistemlerinde eşik değeri tespiti, kullanılan derlemlerde eğitim için ayrılmış olan haberler üzerinde yapılan ön testler sonucu elde edilen değerlerle oluşturulur.

Konu izleme ile ilgili olarak uygulanan yöntemlerde, en önemli iki adım, konu ve haberlerden özellik vektörlerinin çıkarılması ve haber–konu eşleşmelerinde kullanılacak erişim fonksiyonunun seçilmesidir. Özellik vektörlerinin seçilmesi işlemi, kullanılacak erişim fonksiyonu yöntemine doğrudan bağlıdır. Bu kapsamda, konu izleme görevinin gerçekleştirilmesinde, bu görevin başarımını etkileyen en önemli etken, erişim fonksiyonu olarak görülmektedir.

TDT programında, tanımlı görevlerin gerçekleştirilmesinde, erişim fonksiyonu olarak eğiticili (supervised) ve eğiticisiz (unsupervised) olmak üzere iki kategoride yöntemler kullanılmaktadır. Geleneksel bilgi erişim sistemlerinde olduğu gibi, sonucu elde etmek için örnek belgelerin kullanıldığı durumda, eğiticili yöntemler uygulanmakta ve bu yöntemler genel olarak sınıflandırma (classification) olarak adlandırılmaktadır. Diğer

taraftan, sonuç elde etmek için örnek belgelerin kullanılmadığı yöntemler, eğiticisiz olarak çalıştırılmakta ve bu yöntemler genel olarak kümeleme (clustering) olarak adlandırılmaktadır. Geleneksel bilgi erişim sistemleri ve veri madenciliği konusunda, gerek sınıflandırma gerekse kümeleme ile ilgili olarak kullanılan pek çok yöntem bulunmaktadır (Berry ve Castellanos, 2004; Steinbach, Karypis ve Kumar, 2000). Bu yöntemlerin çoğu, aynı zamanda konu izleme görevinin gerçekleştirilmesinde, araştırmacılar tarafından sıkça kullanılmıştır (Dai, Chen, Wang ve Xu, 2010; Xiaowei ve diğerleri, 2008; Li, Lv, Li, ve Shi, 2010a; Li, Lv, Zhouve Shi, 2010b; Diao, Bai ve Yu, 2010).

Özellikle, TDT 2000 toplantılarında, TDT çalışmalarında hikâye bağlantı algılama ve konu izleme görevlerinin önemine vurgu yapılmış ve araştırmacıların bu görevlerin gerçekleştirilmesinde farklı yöntemler kullanmaları önerilmiştir (Allan, 2002). Bu çalışmalarda, konu izleme için tek haber kullanmak ile birden fazla haber kullanmak arasında ciddi bir performans farkı olmadığına dikkat çekilmiş ve sistem başarımının artırılması için, bu farklı yöntemlerin birleşimlerinin test edilmesi konusuna vurgu yapılmıştır (Allan, Lavrenko, Frey ve Khandelwal, 2000).

Diao ve diğerleri (2010), k-en yakın komşu (K-Nearest Neighbor - KNN) algoritmasının varlık isimleri kullanılarak geliştirilmiş bir sürümünü kullanarak, konu izleme görevindeki başarımını test etmişlerdir. Araştırmacılar, uygulanan yöntemin klasik KNN'e göre hem daha başarılı sonuçlar verdiğini, hem de zaman maliyetinin azaldığını rapor etmişlerdir (Diao ve diğerleri, 2010).

Zhang ve diğerleri (2009), konu izleme görevinin başarıyla gerçekleştirilebilmesi için, sadece haberlerde geçen kelime kesişmelerine bakılmasının yeterli olmayacağını, doğru konuları yakalayabilmek için haberlerin anlamsal ilişkilerinin de dikkate alınması gerektiğini belirtmişlerdir. Bu kapsamda, araştırmacılar, haberleri analiz ederek konuları yakalamak için "Gizli Anlam Dizinleme" (Latent Semantic Indexing - LSI) ve haber benzerliklerini belirlemek için de "Destek Vektör Makinaları" (Support Vector Machines - SVM) yöntemini kullanarak testler gerçekleştirmişlerdir. Araştırmacılar, temel yöntem olarak kabul edilen KNN ve SVM yöntemlerine göre, LSI ile SVM'nin

birlikte kullanımının daha başarılı ve etkin sonuçlar verdiğini göstermişlerdir (Zhang ve diğerleri, 2009).

Franz, McCarley, Ward ve Zhu (2001) ise konu algılama (topic detection) görevi için kullanılan eğiticili ve eğiticisiz kümeleme yöntemlerini konu izleme görevinin başarımını test etmek için kullanmışlardır. Araştırmacılar uyguladıkları çift eşik değerli kümeleme yaklaşımını, belge benzerliklerini tespit ederken, temel yöntem olarak kabul ettikleri OKAPI formülünü uygulayarak elde ettikleri sonuçlarla karşılaştırmışlar ve anlamlı bir başarım artışını rapor etmişlerdir (Franz ve diğerleri, 2001).

Yang ve diğerleri (2000), klasik bilgi erişim sistemlerinde, sorgu genişletmek için kullanılan Rocchio yaklaşımını, KNN algoritması ile birlikte kullanarak, konu izleme testlerini gerçekleştirmişler, bu yeni yaklaşımın tek başına KNN kullanılarak elde edilen başarıma göre yaklaşık %71 daha başarılı sonuçlar ürettiğini göstermişlerdir.

Can ve diğerleri (2010) ise; Türkçe bir derlem üzerinde ilk kez gerçekleştirilen çalışmada, statik ve adaptif konu izleme yöntemleri kullanarak, konu izleme başarımını değişik senaryolar üzerinde test etmişlerdir. Araştırmacıların kullandığı statik yöntemde, konular için eğitim haberleri kullanılarak, konu kümeleri başlangıçta bir kez yaratılmakta ve tüm testlerde bu başlangıç kümeleri kullanılmaktadır. Adaptif yöntemde ise, başlangıç kümeleri yaratıldıktan sonra, sisteme ilgili belge ulaştıkça kümeler dinamik olarak güncellenmekte, küme merkezi vektörleri her seferinde yeniden oluşturulmaktadır. Bu çalışmada haber-konu eşleşmelerinde kosinüs benzerliği (cosine similarity) ve kapsayan katsayı tabanlı benzerlik (cover coefficient based similarity) yöntemleri kullanılmıştır. Araştırmacılar, kosinüs benzerliği yönteminde, haberler için en yüksek *tf.idf* değerine sahip 60 kelimeyi seçerek kapsayan katsayı tabanlı benzerlik yöntemine göre daha başarılı sonuçlar elde etmişlerdir. Ayrıca, kullanılan her iki yöntemin, AND mantıksal operatörü ile birleştirilmiş sonuçlarının, kosinüs benzerliği yöntemi ile elde edilen sonuçlardan daha yüksek başarıma sahip olduğu rapor edilmiştir. Bu çalışmadan elde edilen sonuçlar, Türkçe bir derlem üzerinde, konu izleme görevi ile ilgili olarak gerçekleştirilen ilk çalışma olması açısından son derece önemlidir.

Bu çalışmaların yanında, TDT programında konu izleme ile ilgili olarak gerçekleştirilen akademik çalışmalarda, erişim fonksiyonu olarak pek çok farklı modelin kullanıldığı görülmektedir. Gizli Markov modeli tabanlı yöntemler (Yamron, Carp, Gillick, Lowe, ve Van Mulbregt, 1997; Yamron, Carp, Gillick, Lowe ve Van Mulbregt, 1998), olasılık tabanlı yöntemler (Walls, Schwartz, Jin ve Sista,1999), özet çıkarma tabanlı yöntemler (Bun ve Ishizuka, 2001; Bun ve Ishizuka, 2006), sözcüksel zincir (lexical chains) tabanlı yöntemler (Carthy ve Sherwood-Smith, 2002; Carthy ve Smeaton, 2000; Hatch, Stokes ve Carthy, 2000; Chen, Wang ve Liu, 2005), sıradüzensel kümeleme tabanlı yöntemler (Dai ve diğerleri, 2010), anahtar kelime ve cümle çıkarsama tabanlı yöntemler (keyword and keyphrase extraction) (Lee ve Kim, 2008; Wang, Zhang, Ru ve Ma, 2008), dil modeli tabanlı yöntemler (Viermetz, Skubacz, Ziegler ve Seipel, 2008), vektör uzayı modeli tabanlı yöntemler (Xiaowei ve diğerleri, 2008; Li ve diğerleri, 2010a; Li ve diğerleri, 2010b) ve KNN tabanlı yöntemler (Diao, Bai ve Yu, 2010; Köse ve Ahmadlouei, 2013), konu izleme görevi için kullanılan diğer yöntemler olarak karşımıza çıkmaktadır.

Konu izleme sistemlerinin başarımının test edildiği çalışmalarda, genellikle derlem içerisinde eğitim verisi olarak kullanılan konuyla ilgili haberlerin bir kümesi sisteme verilerek, sistemlerin, derlemin kalan kısımları içinde, ilgili diğer haberleri bulabilme yetenekleri test edilir.

Konu izleme görevinin başarımının test edilmesinde, iki önemli konuya dikkat edilmelidir (Allan, 2002; TDT, 2002). Bunlardan birincisi, izleme sistemleri olup, bu sistemlerde her bir konunun izlenecek diğer konulardan bağımsız olarak eğitim ve test işlemlerine sokulması gereklidir. Geliştirilen izleme sistemleri, görevlerini kolaylaştıracak olsa bile, bir konu için diğer konulara ait olan sistem kararlarını kullanamaz. Göz önünde bulundurulması gereken ikinci parametre, konular için sistem tarafından verilen skor değerlerinin normalleştirilme yöntemleridir. Sistemin verdiği kararın ne kadar güvenilir olduğunu gösteren skor değerlerinin, tüm konular için aynı anlamları ifade etmesi zorunludur. Örneğin, bir haber çifti için verilen 0,85 gibi bir skor değeri, başka bir haber çifti için de aynı miktarda kanıt bulunduğunu göstermelidir.

Konu izleme ile ilgili olarak, akademik çalışmalara genel olarak bakıldığında, araştırmaların, hikâye bağlantı algılama görevinde olduğu gibi, ağırlıklı olarak belge gösterimi, erişim fonksiyonu olarak farklı yöntemlerin kullanılması ve kullanılan bu farklı yöntemlerin birleşimlerinin test edilmesi konularında yoğunlaştığı görülmektedir. Özellikle, konu modellerinin oluşturulması ve konuyla ilgili haberlerin tespit edilmeye çalışıldığı farklı katmanlarda, ilgili katmanın doğasına uygun güçlü yöntemlerin kullanılması ile gerçekleştirilen sistemlerdeki başarım artışları dikkat çekicidir.

3. BÖLÜM

YÖNTEM

3.1. BAŞARIM TESTLERİNDE UYGULANAN YÖNTEMLER

Bu bölümde, hikâye bağlantı algılama ve konu izleme görevlerinin başarım testlerinde kullanılan erişim fonksiyonları, test derlemi, test senaryoları ve değerlendirme yöntemleri ile ilgili bilgiler sunulacaktır. Erişim fonksiyonu olarak vektör uzayı modeli ve ilgi modeli, hem hikâye bağlantı algılama hem de konu izleme testlerinde, Canopy ve K-ortalamalar kümeleme algoritmaları ise sadece konu izleme başarım testlerinde kullanılmıştır.

3.1.1. Vektör Uzayı Modeli

Vektör uzayı modeli, klasik bilgi erişim sistemleri tarafından erişim fonksiyonu olarak sıkça kullanılan (Salton ve diğerleri, 1975; Salton, 1989; Frakes ve Baeza, 1992; Schultz ve Liberman, 1999), 1970'lerin başlarında geliştirilmiş olan ve günümüzde halen yoğun olarak kullanılan, oldukça popüler bir yöntemdir. Bu yöntemi kullanan bilgi erişim sistemlerinde, sorgular ve belge koleksiyonunda bulunan her bir belge, koleksiyonda bulunan t_1, t_2 . . . t_n gibi n adet tekil kelimeden oluşan bir vektör gibi gösterilir. Belgenin vektör biçiminde gösterilmesinde kullanılan t_1, t_2 . . .t_n katsayılarının değerleri, ilgili koleksiyon kelimesinin (t_i), belge veya sorgu içerisinde bulunup bulunmamasına ya da kaç kez bulunduğuna göre belirlenir.

t_1, t_2 . . . t_n katsayılarını belirlemek için kullanılan iki farklı yöntem vardır. Bu yöntemlerden birincisi, "*ikili ağırlık (binary weighted) gösterimi*" adı verilen ve ilgili koleksiyon kelimesi belge veya sorgu içerisinde bulunuyorsa t_i değerini 1, bulunmuyorsa 0 olarak kabul eden bir yaklaşım sunar. Bu yöntemde, ilgili koleksiyon kelimesinin belge içerisinde ne kadar sıklıkla kullanıldığı hiçbir önem taşımamaktadır. İkinci yöntem ise "*terim ağırlıklı (term weighted) gösterim*" olarak adlandırılır ve her bir koleksiyon kelimesinin ilgili belge veya sorgu üzerindeki sıklık değerini, vektör katsayıları olarak kullanır.

Bilgi erişim sistemi üzerindeki sorgu ve belgeler, vektör cinsinden ifade edildikten sonra, bunların benzerliklerinin belirlenerek ilgili belgelere, bu benzerliklerin ne kadar güçlü olduğunu ifade eden skor değerlerinin verilmesi gerekir. Vektör uzayı modelinin en büyük avantajı, erişim sistemindeki sorgu ve belgeler arasındaki benzerlik değerinin, sorgu ve belge vektörleri arasındaki uzaklık ölçülerek elde edilebilecek olmasıdır. İki vektör arasındaki uzaklık, bu iki vektörün iç çarpımları alınarak hesaplanabilir. Buna göre, sorgu ve ilgili belge vektörleri, birbirine ne kadar yakınsa, bunların konu olarak birbirine benzer olma olasılıkları da o kadar büyük olacaktır. Böylece, ilgili belgeler sorgu vektörüne yakın olma derecelerine göre derecelendirilebilir.

Sorgu ya da belge vektörleri katsayıları, her bir terimin, bu sorgu ya da belgeleri ifade ederken ne ölçüde önemli olduğunu göstermektedir. Bilgi erişim sistemlerinde bugüne kadar yapılan çalışmalar, bir sözcük bir belge içerisinde ne kadar sık geçerse, o sözcüğün ilgili belgeyi ifade etmekte o kadar etkili olduğunu; öte yandan bir sözcük koleksiyondaki diğer belgeler içerisinde ne kadar sık geçerse, o sözcüğün ayırdedici özelliğinin o kadar az olduğunu göstermektedir. Vektör uzayı modelinde terim ağırlıklarının yukarıda bahsedilen özellikler göz önüne alınarak belirlenmesi genellikle *idf ağırlıklı kosinüs katsayısı* olarak tanımlanır ve *tf.idf (term frequency x inverse document frequency)* olarak gösterilir (Salton ve McGill, 1983).

Bir bilgi erişim sisteminde kullanıcı sorguları ile ilgili belgelerin belirlenmesinde vektör uzayı yönteminin kullanılması iki ön görevin öncelikli olarak yerine getirilmesini gerektirir. İlk görev, özellik seçiminin gerçekleştirilmesidir ve bu adımda kullanıcı sorgusu ve derlemdeki belgeleri ifade eden özelliklerin bir kümesi seçilir (kelimeler ya da kelime gövdeleri). İkinci adımda ise benzerlik ölçümünün sağlanabilmesi için, gerekli bir eşik değeri belirlenir ve bu eşik değerine göre benzerlik kararları verilir. Sorguyla ilgili belgelerin belirlenebilmesi için, sorgu ve belgelerin özellik vektörleri belirlendikten sonra, bu iki vektör arasındaki kosinüs açısı Eşitlik 1'de olduğu gibi hesaplanarak vektörlerin birbiri ile ne kadar benzer oldukları belirlenir (Schultz ve Liberman, 1999).

$$\cos(\theta) = \frac{\vec{A}.\vec{B}}{\|A\|\|B\|} \qquad (1)$$

Bilgi erişim sistemlerinde, vektör uzayı modeli uygulanırken, ikili ağırlıklandırma yerine genellikle *tf.idf* ağırlıklandırması kullanılmaktadır. Daha önce de açıklandığı gibi, bu ağırlıklandırma, bir belge içerisinde çok sık geçen kelimelerin belgeyi ifade etmek için önemli, derlem içerisinde pek çok belgede geçen kelimelerin ise daha az önemli ya da önemsiz olduğu anlamına gelir. Bu ağırlıklandırmada *tf* (*term frequency*), ilgili kelimenin bir belge içerisindeki geçme sıklığını ifade ederken, *idf* (*inverse document frequency*), ilgili kelimenin tüm derlem içerisindeki kullanım sıklığını belirler ve Eşitlik 2'de gösterildiği gibi hesaplanır. Bu formülde *w*, ilgili kelimeyi, *N*, derlemde bulunan toplam belge sayısını ve *df(w)* ise *w* kelimesinin bulunduğu toplam belge sayısını ifade etmektedir. Sonuç olarak, belge vektöründe ifade edilen kelimelerin ağırlıkları, *tf.idf* çarpımı ile bulunarak, ilgili belgeyi ifade etmek için gerçekten önemli olan kelimelerin ağırlıkları artırılmış olur.

$$idf(w) = \log_{10}\left(\frac{N}{df(w)}\right) \qquad (2)$$

Sonuç olarak, *idf* ağırlıklı olarak yapılan benzerlik ölçümleri (*sim(a,b)*) Eşitlik 3'te gösterildiği gibi hesaplanır. Bu ölçüm, iki farklı belgeden oluşturulan vektörlerin arasındaki kosinüs açısına göre benzerlik oranlarını belirlemektedir. Eşitlik 3'te kullanılan $tf_a(w)$, *w* kelimesinin *a* belgesi içerisindeki sıklığı, $tf_b(w)$, *w* kelimesinin *b* belgesi içerisindeki sıklığı ve *idf(w)* de *w* belgesinin derlem içerisindeki sıklığını ifade etmektedir.

$$sim(a,b) = \frac{\sum_{w=1}^{n} tf_a(w).tf_b(w).idf(w)}{\sqrt{\sum_{w=1}^{n} tf_a^2(w)}.\sqrt{\sum_{w=1}^{n} tf_b^2(w)}} \qquad (3)$$

3.1.2. İlgi Modeli

Klasik bilgi erişim sistemlerinde, kullanıcı sorguları ve bu sorgularla ne kadar ilgili oldukları bilinmeyen, büyük boyutta belge kümeleri bulunur ve bilgi erişim sistemlerinin bu kümelerdeki belgeleri, sorgularla ilgili olma olasılıklarına göre derecelendirmeleri beklenir. Bu derecelendirme, çoğunlukla sorgu kelimeleri ile dizin terimlerinin çakışma oranlarına göre gerçekleştirilmektedir. Bu kapsamda gerçekleştirilen çalışmalardaki amaç, sorgu ve derlemdeki belgelerde bulunan kelimelerin çakışma olasılıklarını artırmaktır. Bu konuda, vektör uzayı modelinde sorgu genişletme teknikleri kullanılırken (Xu ve Croft, 2000), istatistiksel yöntemlerde, geçmiş modellerden yararlanılarak sorguyu genişletmek için ilgili olduğu düşünülen kelimeler doğrudan kullanılır (Yamron, Knecht ve Mulbregt, 2000; Leek ve diğerleri, 2002).

Bu kapsamda, bilgi erişim sistemlerinde yoğun olarak kullanılan bir diğer yöntem de, dil modelidir (language model). Dil modeli, sorguların belgelerle ilgili olma olasılıklarını kestirerek belgeleri derecelendirmeyi amaçlar ve bir belgeyi $w_1, w_2, \ldots, w_n$ kelimelerinden oluşan bir kelimeler kümesi gibi ifade ederek kullanır (Song ve Croft, 1999; Ponte ve Croft, 1998; Lavrenko ve Croft, 2001). Dil modeli ve bu modelin bir türevi olan ilgi modeli (relevance model), klasik bilgi erişim sistemlerinin dışında, TDT programında, hikâye bağlantı algılama görevinin gerçekleştirilmesinde yoğun olarak kullanılmıştır (Berger ve Lafferty, 1999, Miller ve diğerleri, 1999; Song ve Croft, 1999; Lavrenko ve Croft, 2001).

İlgi modeli, dil modelinin uygulanması için gerekli olan eğitim verilerinin bulunmadığı ortamlarda, olasılıkların kestirilmesi için yeni bir yaklaşım sunmaktadır. Lavrenko ve Croft (2001) ilgi modelini, "Bir sorgu ile ilgili bir belge içerisinde, *w* kelimesinin bulunma olasılığını ifade eden ve *R*'nin sorguyla ilgili belgelerin kümesini gösterdiği bir evrende, $P(w|R)$ koşullu olasılığının kestirilmesini sağlayan mekanizma" olarak tanımlamışlardır.

Dil modeli üzerine geliştirilen ilgi modeli gibi tüm yaklaşımlar, kullanıcının bilgi ihtiyacını ifade ettiği sorguları genişleterek, koleksiyon içerisinde daha fazla sayıda

ilgili belgeye ulaşmayı amaçlamaktadır. Bu bağlamda ilgi modeli, elimizde sadece kullanıcı sorgusu ve içerisinde çok sayıda ilgili ve ilgisiz belge barındıran, büyük bir koleksiyonun bulunduğu bir durumda, kullanıcı sorgusunun nasıl genişletilebileceğini göstermektedir. Bu yaklaşıma göre, basit olarak, ilgi modeli kullanıcı sorgusunu kullanarak koleksiyon içerisinde bu sorguda kullanılan kelimelerin tamamını ya da belirli bir kısmını içeren ilgili belgeleri getirmektedir. Daha sonra, getirilen bu belgelerde bulunan her bir kelime, getirilen ilgili küme içerisinde bulunma olasılıklarına göre derecelendirilmekte ve en yüksek olasılık değerine sahip *n* adet kelime alınmaktadır. Bu *n* adet ilgili kelime kullanılarak sorgu genişletilmekte ve oluşturulan bu model ilgi modeli olarak adlandırılmaktadır.

Klasik bilgi erişim sistemlerinde kullanılan kullanıcı sorgusunun yerini, TDT çalışmalarında karşılaştırılacak olan belgeler almaktadır. Buna göre, karşılaştırılacak S_1 ve S_2 belgeleri için, ilgi modelleri oluşturularak daha sonra bu modellerin karşılaştırılması sağlanabilir. Bu noktada, belirli bir belge için oluşturulan ilgi modelinin, o belgenin içerdiği konuyu yakalaması beklendiği için artık ilgi modeli yerine konu modeli (topic model) kavramı kullanılacaktır.

Konu modelini oluşturmak için, öncelikle *S* belgesi, ilgi modelinde kullanılan sorgu gibi düşünülür. Buna göre, $S = q_1, q_2 \ldots q_k$'dan oluşan sorgu kullanılarak eğitim belgeleri içinde, bu sorguyla ilgili olan belgelere ulaşılır. Bu noktada, erişilen belgelerden, sorguyla en çok ilgili olan *m* adet belgeye ihtiyaç duyulmaktadır. Bu ayrımı gerçekleştirmek için, erişilen her bir eğitim belgesi (D), $P(D|S)$ ya da $P(D|q_1 \ldots q_k)$ olasılığına göre derecelendirilir (Lavrenko ve diğerleri, 2002). Her bir belge için, bu olasılık değerinin hesaplanarak, ilgili belgeye bir sıra değeri vermek için Eşitlik 4 kullanılabilir.

$$P(D \mid q_1 \ldots q_k) = \frac{P(q_1 \ldots q_k \mid D) P(D)}{P(q_1 \ldots q_k)} \qquad (4)$$

Bu eşitlikte, $P(D)$ ve $P(q_1 \ldots q_k)$ olasılıkları erişilen belge içerisinde sabit olduğundan, sonuç üzerinde sadece $P(q_1 \ldots q_k \mid D)$ olasılığı etkili olacaktır. Bu noktada, sorgu

kelimelerinin birbirinden bağımsız olarak seçildikleri varsayılarak[1] $P(q_1..q_k \mid D)$ olasılığı Eşitlik 5'te gösterildiği gibi hesaplanabilir.

$$P(q_1...q_k \mid D) = \prod_{i=1}^{k} P(q_i \mid D) \qquad (5)$$

Eşitlik 5'te gösterilen ve sorgu kelimelerinin sayısını gösteren *k*, eğer birkaç kelimeden büyükse (ki öyle olması beklenmektedir), sonuç olasılığının sıfıra doğru gideceğini söylemek yanlış olmaz ($P(q_i \mid D) < 1$ olduğu hatırlanmalıdır). Sonuç olarak, eğitim derleminden erişilen belgeler için verilen sıra değerleri, genellikle kayan noktalı çok küçük sayılar olacak ve sadece çok ilgili olan belgeler anlamlı sıra değerlerine sahip olacaktır. İlgili tüm eğitim belgelerine anlamlı sıra değerleri vermek için $P(D \mid q_1...q_k)$ olasılığının *k.* kökü alınarak bu problem ortadan kaldırılır.

S hikâyesi bir sorgu olarak kullanılarak eğitim derlemi içerisinde ulaşılan *m* adet ilgili belgeye, R_q ilgi kümesi adı verilir. Bu noktada, erişilen belgeler içerisinde bulunan her bir kelime, R_q ilgi kümesinde bulunma olasılıklarına göre derecelendirilir ve en çok ilgili olan *k* adet kelime *S* hikâyesi için konu modelini oluşturur. İlgili belge içerisinde bulunan *w* kelimesinin R_q ilgi kümesi içerisinde bulunma olasılığı Eşitlik 6'da olduğu gibi ifade edilebilir.

$$P(w \mid R_q) = \sum_{D \in R_q} P(w \mid D) P(D \mid Q) \qquad (6)$$

Pek çok dil modeli yaklaşımı $P(w|D)$ olasılığını hesaplamak için, kelimenin koleksiyon içerisinde bulunma olasılığını ve doğrusal aradeğerleme yapılan maksimum benzerlik (maximum likelihood) kestirmesini kullanır (Eşitlik 7).

$$P(w \mid D) = \lambda P_{ml}(w \mid D) + (1-\lambda) P_{bg}(w) = \lambda \frac{tf_{w,D}}{|D|} + (1-\lambda) \frac{cf_w}{coll.size} \qquad (7)$$

[1] Bu bağımsızlık varsayımı bazı araştırmacılar tarafından çok doğru bulunmamakla birlikte, hesaplamayı kolaylaştırması tercih sebebi olmuştur (Robertson, Maron ve Cooper, 1983).

Yukarıda anlatılan yaklaşım kullanılarak, her bir belge için konu modelleri oluşturulduktan sonra, karşılaştırılan belgelerin konu benzerliklerinin belirlenmesinde, bu iki belge için oluşturulan konu modellerinin karşılaştırılması gerekir. Verilen S_1 ve S_2 belgeleri için bunların konu modellerinin sırasıyla M_1 ve M_2 olduklarını kabul edelim. İstatistiksel yöntemlerle bu iki konu modelinin benzerliklerinin karşılaştırılması için, hem $P(S_1|M_2)$ hem de $P(S_2|M_1)$ olasılıklarının kestirilmesi gerekir. Oysa elimizde benzer miktardaki veriler üzerinde kestirilmiş olan iki olasılık dağılımı modeli varsa, bunların doğrudan karşılaştırılması mümkündür. Bu noktada, *Kullback-Leibler (KL)* uzaklığı yöntemi, iki farklı olasılık dağılımının birbirinden ne kadar farklı olduklarını belirlemek için standart bir yol sunar (Eşitlik 8).

$$D(M_1||M_2) = \sum_w P(w|M_1) log \frac{P(w|M_1)}{P(w|M_2)} \qquad (8)$$

KL uzaklığı, asimetrik bir yaklaşım sunar. Yani, M_1 modelinin M_2 modelinden ne kadar farklı olduğunu bulduğumuzda, aynı zamanda M_2 modelinin M_1 modelinden ne kadar farklı olduğunu da bulmuş olmayız (Lavrenko ve diğerleri, 2002). Oysa konuyla ilgili kaynakların benzerliklerinin belirlenebilmesi için simetrik bir ölçüm gereklidir. Bu problemi çözmek için her iki yöndeki KL uzaklıkları hesaplanarak, bulunan değerlerin toplanması ile $(D(M_1||M_2) + D(M_2||M_1))$ elde edilen simetrik hesaplama kullanılır. KL uzaklığı, iki olasılık dağılımı arasındaki farkları gösterdiği için, benzerliklerin belirlenmesinde yukarıdaki eşitliğin negatif değeri kullanılır. Bu hesaplama biçimi, kabul edilebilir bir yaklaşım olarak görülmesine rağmen bazı problemlere sahiptir. Eğer üretilen modeller çok belirsizse ya da modeller genel dile çok benziyorsa, bu modellerin karşılaştırılmasından anlamlı sonuçlar elde etmek mümkün olmayacaktır. Bu problemi çözmek için modellerin karşılaştırılması aşamasında, ilgili olasılık dağılımı ve genel dil modeli arasındaki KL uzaklığı değeri alınarak, modellerin açık olmaları sağlanmalıdır (Lavrenko ve Croft, 2001). Böylece olasılık dağılımı genel dil modeline benzemediğinde, anlamlı sonuçlar üretilecektir. Sonuç olarak, elimizdeki konuların benzerliklerini belirlemek için kullanacağımız asimetrik yaklaşım $[-D(M_1 \,||\, M_2) + Clarity(M_1)]$ biçiminde olacaktır. Bu formül, M_1 modelinin genel dilden ne kadar farklı olduğunu belirleyen $Clarity(M_1)$ parametresi ile genişletilmiş olan, M_2 ve M_1 modellerinin ne kadar benzer olduklarını gösteren yaklaşımı oluşturmaktadır. Bazı

küçük değişikliklerle sözü edilen benzerlik yaklaşımı Eşitlik 9'da olduğu gibi ifade edilebilir.

$$\sum_{w} P(w|M_1)\log\frac{P(w|M_2)}{P(w|GE)} \quad (9)$$

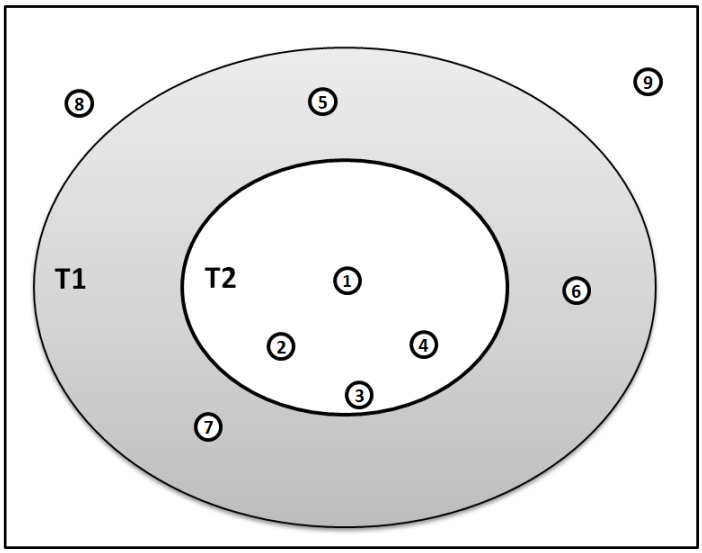
Bu eşitlikte kullanılan açıklık (*clarity*) parametresinin, belge erişiminde kullanılan *idf* parametresi ile benzer bir rol üstlendiğine dikkat edilmelidir. Yani, genel dil içerisinde çok sık kullanılan kelimelerin, bir konuyu ifade ederken daha düşük ağırlıklara sahip olması sağlanmaktadır. Sonuç eşitliğine ulaşmak için, M_1 ve M_2 modellerinin yerleri değiştirilerek ikinci bir hesaplama yapılır ve benzerlik değeri elde edilir.

3.1.3. Canopy Kümeleme Algoritması

Kümeleme; en genel tanımı ile dağınık bir veri kümesi içerisinde öznitelikleri birbirine benzeyen elemanların bir araya getirilerek gruplandırılması işlemi olarak tanımlanabilir (Han, 1996). Kümeleme işlemi sonucunda ortaya çıkan bir küme, kendi içerisinde benzer özniteliklere sahip elemanlar barındıran, ancak elemanları diğer kümelerin elemanlarının özniteliklerinden farklı olan bir grup olarak ifade edilebilir (Larose, 2005).

Kümeleme işleminde, işlemin doğası gereği, çok sayıda karmaşık işlemin tekrarlı olarak yapılmasını gerektiren teknikler kullanılmaktadır. Bu kapsamda, büyük boyutlu veriler üzerinde kümeleme işlemi, popüler kümeleme yöntemlerinin kullanıldığı yinelemeli (iteratif) yaklaşımların karmaşıklığından kaynaklanan nedenlerden ötürü, ciddi bir zaman maliyetine sahiptir. Bu nedenle, büyük veri kümelerini doğruluk kıstasını ikinci plana atarak hızlı bir biçimde gruplandırmak, önemli bir ihtiyaç olarak ortaya çıkmaktadır. Bu probleme çözüm olarak geliştirilen Canopy, son derece basit, hızlı ve doğru sonuçlar üreten bir kümeleme algoritmasıdır (McCallum, Nigam ve Ungar, 2000). Canopy, bu yapısı ile genellikle daha karmaşık kümeleme algoritmaları kullanılmadan önce, büyük veri kümesinin basit ancak hızlı bir biçimde gruplanması amacıyla kullanılmaktadır.

Canopy algoritması, Şekil 3'te gösterildiği gibi, çok boyutlu bir uzayda kümelenecek veri noktaları ile T_1 ve T_2'den oluşan ($T_1 > T_2$) iki eşik değerle kümeleme işlemine başlar. Öncelikle kümelenecek noktalardan rastgele bir nokta başlangıç küme merkezi olarak seçilir (şekilde 1 no'lu nokta). Sonraki aşamada hızlı bir mesafe belirleme formülü uygulanarak, merkez noktasının diğer tüm noktalara olan uzaklığı hesaplanır. Bu hesaplama sonunda, bazı noktalar küme merkezine olan T_2 uzaklığında kalırken (şekilde 2, 3 ve 4 no'lu noktalar), bazı noktalar T_2 uzaklığının dışında T_1 uzaklığının içinde (şekilde 5, 6 ve 7 no'lu noktalar) ve bazı noktalarda T_1 (şekilde 8 ve 9 no'lu noktalar) uzaklığının dışında kalır.



Şekil 3. Canopy kümeleme algoritması başlangıç durumu

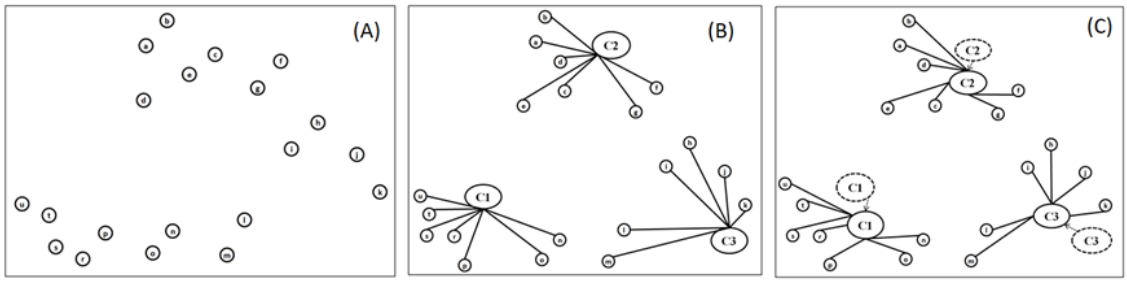
Bu başlangıç hesaplamasında belirlenen küme merkezine, T_2 uzaklığında bulunan bütün noktalar, bu kümenin bir elemanı olarak kabul edilir. Bu noktalar başka bir kümenin merkezi olamaz ve oluşan bu kümeye "canopy" adı verilir. Bunun yanında küme merkezine T_2 mesafesinden daha uzak olan ancak T_1 mesafesi içerisinde kalan diğer tüm noktalar, yeni canopy merkez noktası adayları olarak bir listeye yazılır ve bir aday küme merkez noktaları listesi oluşturulur. Sonraki aşamada, aday listede bulunan her bir nokta, yeni canopy merkezleri olarak seçilerek, küme elemanlarını belirleme işlemi, aday listedeki tüm noktalar bitene kadar tekrarlanır. Böylece, algoritmanın tüm yinelemeleri tamamlandığında, kümelerin merkezi noktaları ve bu kümelere ait olan elemanlar belirlenmiş olur (McCallum ve diğerleri, 2000).

Canopy algoritması sayesinde, başlangıçta oluşturulacak küme sayısının belirlenmesine gerek kalmadan veri hızlı bir biçimde gruplandırılabilmektedir. Genellikle canopy tarafından üretilen kümeler, bir başlangıç kestirmesi olarak kullanılmakta ve bir sonraki aşamada k-ortalamalar, KNN ya da hiyerarşik bir kümeleme algoritması kullanılarak sonuç kümeleri elde edilmektedir.

3.1.4. K-Ortalamalar Kümeleme Algoritması

1967 yılında geliştirilen k-ortalamalar algoritması, en eski kümeleme algoritmalarındandır (MacQueen, 1967). K-ortalamalar algoritması öznitelikleri farklı elemanlardan oluşan bir veri yığınında *N* adet elemanı *K* adet kümeye bölmeyi hedefler. Bu kapsamda, canopy algoritmasından farklı olarak, başlangıçta ulaşılmak istenen küme sayısını ifade eden *K* değerinin belirlenmesi gerekmektedir. K-ortalamalar algoritması, diğer kümeleme yaklaşımlarında olduğu gibi, aynı kümelerdeki elemanların öznitelik benzerliklerinin yüksek, kümeler arasındaki elemanların öznitelik benzerliklerinin ise düşük olmasını hedefler (Han ve Kamber, 2006).

K-ortalamalar yönteminde, Şekil 4(A)'da gösterildiği gibi, kümelenecek olan veriler ve bu verilerin kümelenmesi istenen küme sayısı (şekilde küme sayısı 3 olarak seçilmiştir) bilgisi ile algoritma yürütülmeye başlanır.



Şekil 4. K-ortalamalar algoritması adımları

Algoritmada ilk aşamada, veri düzlemi üzerinde küme sayısı kadar rastgele nokta, oluşturulacak olan kümelerin merkezlerini (centroid) belirlemek üzere seçilir (şekilde C_1, C_2 ve C_3). Sonraki aşamada, düzlem üzerindeki her bir noktanın, belirlenen küme merkezlerine uzaklıkları, bir uzaklık belirleme formülü (Khachumov, 2012) kullanılarak hesaplanır ve her bir nokta, Şekil 4(B)'de gösterildiği gibi, kendisine en yakın küme

merkezine bağlanır. Bu aşamada, kümelere bağlanan noktaların değerleri kullanılarak, bu değerlerin ortalamaları alınmak sureti ile kümelerin merkez noktaları yeniden hesaplanır (Xu veWunsch, 2005). Bu hesaplama sonunda, Şekil 4(C)'de gösterildiği gibi küme merkezleri yeni konumlarına yerleşir. Küme merkezleri yeniden belirlendikten sonra, düzlem üzerinde yer alan tüm noktaların yeni küme merkezlerine uzaklıkları tekrar hesaplanır ve her bir nokta kendisine en yakın küme merkezine bağlanır. Bu işlem, düzlem üzerindeki noktaların ait oldukları küme merkezleri değişmeyene kadar devam eder ve algoritmanın çalışması sonlandığında düzlemdeki tüm noktalar bir kümeye atanmış olur.

K-ortalamalar, merkez noktanın kümeyi temsil etmesi yaklaşımına dayalı bir yöntemdir ve her verinin sadece bir kümeye ait olabilmesine izin verir. Bu nedenle, keskin bir kümeleme algoritmasıdır (Han ve Kamber, 2006). Bunun yanında başlangıçta oluşturulması gereken küme sayısını kestirmenin zor olması, başlangıç merkez vektörlerini belirleme zorluğu ve algoritmada çevrimlerin ne zaman sonlanacağının başlangıçta bilinmemesi bu yöntemin tartışmalı yönleri olarak kabul edilmektedir (Pena, Lozano ve Larranaga, 1999).

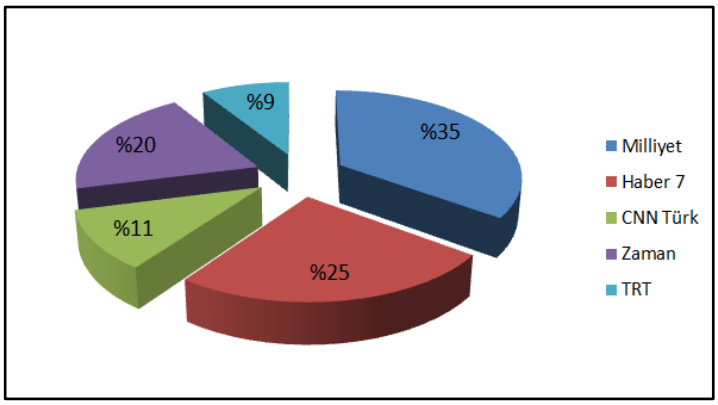
3.2. TEST DERLEMİ

Bu alanda uluslararası akademik çalışmalarda, TDT programı için geliştirilmiş olan TDT Corpora[2] kullanılmaktadır. Ancak, bu çalışmanın kapsamının Türkçe belgeler olması nedeniyle, TDT derleminin kullanılması, tez çalışmasında uygulanacak yöntemlerin hedeflediği alan için uygun değildir. Bu tez çalışmasında Bilkent Üniversitesinde geliştirilmiş olan ve yapısal olarak TDT derleminin özelliklerini içeren, tamamen Türkçe haberlerden oluşan BilCol-2005 derlemi (Can ve diğerleri, 2007) kullanılmıştır.

BilCol-2005 derlemi; CNN Türk, Haber7, Milliyet Gazetesi, TRT ve Zaman gazetelerinin web portallarından 1 Ocak 2005 ve 31 Aralık 2005 tarihleri arasında toplanmış olan toplam 209.305 adet haberden oluşmaktadır (Can ve diğerleri, 2010).

[2] http://www.itl.nist.gov/iad/mig/tests/tdt/resources.html

BilCol-2005 derlemini oluşturan haber kaynakları ve bu haber kaynaklarından derlem içerisinde yer alan haber sayıları Şekil 5'te sunulmuştur.



Şekil 5. BilCol-2005 derlemi kaynağa göre haber dağılımları

3.3. TEST SENARYOLARI

Bu kısımda; hikâye bağlantı algılama ve konu izleme görevlerinin, BilCol-2005 derlemi üzerinde gerçekleştirilen başarım testlerinde uygulanan yöntemlerin aşamaları verilmektedir. Uygulanan tüm yöntemlerde, haberler kullanılmadan önce, bir ön işlemden geçirilmiştir. Bu ön işlem esnasında, tüm karakterler küçük harfe çevrilmiş, noktalama ve özel işaretler ayıklanmış ve Türkçe durma kelimeleri metinlerden çıkarılmıştır. Ancak, metinler üzerinde herhangi bir gövdeleme algoritması çalıştırılmamış ve haberleri oluşturan kelimeler olduğu gibi kullanılmıştır.

Daha önce belirtildiği gibi, BilCol-2005 derleminde bulunan toplam 209.305 adet haberden sadece 5883 tanesi konu başlıklarına (80 konu) göre etiketlenmiş olup, kalan 203.422 adet haberin, aslında hangi konularla ilgili oldukları tam olarak bilinmemektedir. Ancak, ilgili akademik çalışmalarda (Can ve diğerleri, 2010) etiketlenmemiş olan haberler, belirlenmiş olan 80 konu başlığından farklı olarak kabul edildiği için bu çalışma da bu kabul üzerine oturtulmuştur. Bu kapsamda hikâye bağlantı algılama testleri, BilCol-2005 derlemindeki tüm haberler (209.305 adet haber), konu izleme testleri ise konu başlıkları belirlenmiş olan 5883 haber üzerinde gerçekleştirilmiştir.

Başarım testlerini gerçekleştirmek için seçilen haberler iki gruba ayrılmış ve ilk grup eğitim, ikinci grup ise test için kullanılmıştır. Her bir konu başlığında, var olan belge sayısının üçte biri eğitim, üçte ikisi de test belgesi olarak kabul edilmiştir. Tarih sırasına göre derlemdeki ilk N belge eğitim, kalanlar ise test belgesi olarak kullanılmış olup, derlem üzerinde eğitim belgeleri olarak seçilen haberler, test aşamasında uygulanacak olan ilgililik eşik değerini belirlemek amacıyla kullanılmıştır.

Derlem üzerinde her bir yöntem için başarım testleri gerçekleştirilirken, her bir sorgu için ikili sınıflandırma tabloları yaratılmış, tüm sorgular yürütüldükten sonra bu tablolar birleştirilerek, mikro ortalama yöntemi kullanılarak anma, duyarlık ve f-ölçü değerleri hesaplanmıştır.[3]

3.3.1. Hikâye Bağlantı Algılama Test Senaryoları

Hikâye bağlantı algılama testlerinde, vektör uzayı modeli ve ilgi modeli yöntemleri başarım testleri uygulanmış olup testler esnasında uygulanan senaryolar aşağıdaki gibi gerçekleşmiştir.

A. Her bir konu ile ilgili olarak eğitim belgeleri belirlendikten sonra uygun eşik değerinin seçilmesi işlemi şu şekilde gerçekleştirilmiştir:
   - **a.** Öncelikle derlemde bulunan 209.305 belgenin üçte biri (69.768) eğitim belgesi olarak belirlenmiş ve dizinlenmiştir.
   - **b.** İlgililik değerlendirmesi yapılmış olan 5883 belgenin üçte biri olan 1961 belge, eğitim için sorgu olarak kabul edilmiştir.
   - **c.** Her bir sorgu derleme gönderilmiş, ilgi modeli ve vektör uzayı modeli kullanılarak üretilen sorgu-belge eşleşme skorları belirlenmiştir.
   - **d.** Belirlenen tüm bu skor değerleri içerisinden, sorgunun ilgili olduğu bilinen belgeler için üretilen skor değerleri belirlenerek, ilgili sorgu-belge eşleşmeleri için ortalama skor değeri, başlangıç eşiği olarak kabul edilmiştir.
   - **e.** Bu başlangıç eşiğine göre, her bir konu için anma/duyarlık değerleri hesaplanmıştır.

[3] Bu ölçüler aşağıda (3.5) tanımlanmaktadır.

f. Sonraki aşamada, eşik değeri belirli oranda azaltılıp-artırılarak anma/duyarlık değerleri her bir eşik için tekrar hesaplanmıştır.

g. Anma ve duyarlığın birlikte en yüksek oldukları (ya da birbirlerine en yakın oldukları) değer, sistemin kesin eşik değeri olarak belirlenmiş ve sistem testleri bu değere göre gerçekleştirilmiştir.

B. Kesin eşik değeri belirlendikten sonra, test derlemi üzerindeki değerlendirmeler aşağıdaki gibi gerçekleştirilmiştir:

a. Derlemde bulunan 209.305 belgenin üçte ikisi (139.536), test belgesi olarak belirlenmiş ve dizinlenmiştir.

b. İlgililik değerlendirmesi yapılmış olan 5883 belgenin üçte ikisi olan 3922 belge, test için sorgu olarak kabul edilmiştir.

c. Her bir sorgu derleme gönderilmiş, ilgi modeli ve vektör uzayı modeli kullanılarak üretilen sorgu-belge eşleşme skorları belirlenmiştir.

d. Üretilen skor, belirlenen eşik değerine eşit ya da üzerinde ise belge ilgili, değilse ilgisiz olarak kabul edilmiştir.

e. Elde edilen sonuçlara göre, yöntemin başarımı mikro ortalama yöntemi kullanılarak hesaplanmıştır.

C. Vektör Uzayı Modeli ve İlgi Modeli için yukarıda belirlenen her bir aşama, belgeleri ifade etmek için seçilen terim sayısına göre (1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100, 125, 150, 175, 200, 225, 250, 275, 300, 400, 500 ve 1000 terim için) tekrarlanmıştır.

3.3.2. Konu İzleme Test Senaryoları

Konu izleme testlerinde; k-ortalamalar, vektör uzayı modeli ve ilgi modeli olmak üzere üç farklı yöntem uygulanmış olup, her bir yöntemle ilgili olarak farklı senaryolar yürütüldüğü için, yöntemler için uygulanan senaryolar aşağıda farklı başlıklar olarak sunulmuştur. Konu izleme başarımı belirlenirken, testler ilgililik değerlendirmesi yapılmış olan 80 konu başlığındaki 5883 adet haber üzerinden yapılmış, her bir konuda var olan belge sayısının üçte biri eğitim (1961 adet), üçte ikisi de test belgesi (3922

adet) olarak kabul edilmiştir. Her bir konu kümesini yaratmak için tarih sırasına göre ilk 4 belge kullanılmıştır.

3.3.2.1. K-Ortalamalar Yöntemi Test Senaryosu

A. Her bir konu ile ilgili olarak, konu modellerini oluşturmak ve gerekli parametreleri belirlemek için aşağıdaki adımlar gerçekleştirilmiştir:

- **a.** Metin kümeleme işleminde k-ortalamalar algoritması kullanılmıştır.
- **b.** K-ortalamalar algoritmasının yürütülmesi için gerekli başlangıç merkez vektörleri (centroid), Canopy algoritması kullanılarak belirlenmiştir.
- **c.** Başlangıç noktalarına göre k-ortalamalar algoritması yürütülerek, her bir konu için konu kümeleri yaratılmıştır.
- **d.** Her bir konu kümesi için küme merkezleri (centroids) ve her bir kümeye ait olan belge vektörleri belirlenmiştir.
- **e.** Her bir konu merkezi vektörü ile o konuya ait olan belge vektörleri arasındaki uzaklıklar "*Cosine Similarity*" yöntemi ile belirlenmiştir.
- **f.** Eşik değeri belirlenirken, eğitim kümesindeki belgeler sorgu olarak kullanılmış ve en yüksek anma/duyarlık değerindeki eşik değeri, sistem eşiği olarak kabul edilmiştir.

B. Konular için eşik değerleri ve konu kümeleri belirlendikten sonra, test derlemi üzerindeki değerlendirmeler aşağıdaki gibi gerçekleştirilmiştir:

- **a.** 3922 belge test belgesi olarak kullanılmış ve bu belgeler sorgu olarak kabul edilmiştir.
- **b.** Her bir sorgu, daha önce konu modelleri oluşturulmuş olan kümelerle kosinüs benzerliği yöntemi kullanılarak karşılaştırılmıştır.
- **c.** Bu karşılaştırmalar sonucunda elde edilen uzaklık eğitim aşamasında belirlenen eşik değerden düşük ya da bu değere eşitse belge konuyla ilgili, değilse ilgisiz olarak kabul edilmiştir.
- **d.** Elde edilen sonuçlara göre yöntemin başarımı, mikro ortalama yöntemi kullanılarak hesaplanmıştır.

3.3.2.2. Vektör Uzayı Modeli Test Senaryosu

A. 1961 belge, eğitim belgesi olarak kabul edilmiş ve daha önceki yöntemlerde uygulanan eşik bulma yöntemi ile eşik değerler tespit edilmiştir.

B. Test belgesi olarak 3922 belge belirlenmiş ve bu belgeler sorgu olarak kabul edilmiştir.

C. Her bir sorgu ilgili konuyu temsil eden dört eğitim belgesi ile Vektör Uzayı yöntemi kullanılarak karşılaştırılmıştır.

D. Bu karşılaştırmalar sonucunda elde edilen uzaklık eğitim aşamasında belirlenen eşik değerden düşük ya da bu değere eşitse belge konuyla ilgili, değilse ilgisiz kabul edilmiştir.

E. Elde edilen sonuçlara göre, yöntemin başarımı mikro ortalama yöntemi kullanılarak hesaplanmıştır.

3.3.2.3. İlgi Modeli Test Senaryosu

A. 1961 belge, eğitim belgesi olarak kabul edilmiş ve daha önceki yöntemlerde uygulanan eşik bulma yöntemi ile eşik değerler tespit edilmiştir.

B. Test belgesi olarak 3922 belge belirlenmiş ve bu belgeler sorgu olarak kabul edilmiştir.

C. Her bir sorgu, ilgili konuyu temsil eden dört eğitim belgesi ile ilgi modeli kullanılarak karşılaştırılmıştır.

D. Bu karşılaştırmalar sonucunda elde edilen uzaklık, eğitim aşamasında belirlenen eşik değerden düşük ya da bu değere eşitse belge konuyla ilgili, değilse ilgisiz kabul edilmiştir.

E. Elde edilen sonuçlara göre, yöntemin başarımı mikro ortalama yöntemi kullanılarak hesaplanmıştır.

3.4. KULLANILAN ARAÇLAR

Bu tez çalışması kapsamında, kullanılan yöntemlerin başarım testlerini gerçekleştirmek amacıyla, vektör uzayı modeli için Lucene, ilgi modeli için Lemur ve kümeleme için Mahout açık kaynak kodlu yazılımlarından yararlanılmıştır.

Lucene, Doug Cutting tarafından geliştirilmiş ve Apache tarafından desteklenen açık kaynak kodlu, java tabanlı, yüksek performanslı ve ölçeklenebilir bir bilgi erişim kütüphanesidir. Bu kütüphane sayesinde, uygulamalara, istenilen kapsamda tam metin (full text), dizinleme (indexing) ve arama (searching) yetenekleri eklenebilmektedir. Lucene, belgelerin dizinlenmesi esnasında ters dizin kütüklerini (inverted index) kullanırken, sorgu-belge eşleştirmelerinde vektör uzayı yönteminden yararlanmaktadır. Tez çalışmasında, vektör uzayı modelinin uygulanmasında, bu alanda oldukça güçlü olarak kabul edilen ve pek çok bilimsel araştırmada kullanılan Lucene kütüphanesi kullanılmıştır.

Lemur, Massachusetts Üniversitesinden Center for Intelligent Information Retrieval (CIIR) çalışma grubu, Amherst Üniversitesi ve Carnegie Mellon Üniversitesinden Language Technologies Institute (LTI) çalışma gruplarının ortak çalışmaları sonucu üretilmiş olan bir bilgi erişim ve metin analiz kütüphanesidir. Lemur projesi, büyük boyutlu metin kümelerinin ön işlemlerinin (preprocessing) yapılması, dizinlenmesi ve bu dizinler üzerinde sorgular yürütülmesi için gerekli araçları içermektedir. Lemur, sorgu-belge eşleştirmelerinde dil modelinin gelişmiş bir sürümü olarak kullanılan ilgi modelini (relevance model) kullanmaktadır. Özellikle TDT programında pek çok akademik çalışmada kullanılmış olan bu kütüphane, ilgi modeli ile ilgili yöntemin test edilmesinde kullanılmıştır.

Mahout[4] ise Apache tarafından desteklenen ve büyük boyutlu veriler üzerinde, makine öğrenme algoritmalarının etkin bir biçimde yürütülmesini sağlamayı hedefleyen bir kütüphanedir. Mahout içerisinde ağırlıklı olarak kümeleme ve sınıflama konularında akademik çalışmalarda yoğun olarak kullanılan pek çok algoritmanın gerçekleştirimi bulunmaktadır. Bu tez çalışması kapsamında da, kümeleme testlerinde kullanılan k-

[4] https://mahout.apache.org/

ortalamalar ve canopy algoritmaları Mahout Kütüphanesi kullanılarak gerçekleştirilmiştir.

3.5. PERFORMANS DEĞERLENDİRME

Geçmişten günümüze bilgi erişim sistemleri üzerinde gerçekleştirilen ve bu çalışma kapsamında genel olarak incelenen tüm bu yöntemlerdeki ortak amaç; sistemlerin başarım ya da etkinliklerinin artırılmasıdır. Bilgi erişim sistemlerinin etkinliği (ya da başarımı) genellikle anma ve duyarlık değerleri ile belirlenir. Bu değerlerin hesaplanmasında, ikili sınıflama tablosu kullanılır (Tablo 1). Tabloda, "*a*" sistem tarafından erişilen ve kullanıcının ilgili (relevant) bulduğu belge sayısını, "*b*" sistem tarafından erişilen ancak kullanıcının ilgisiz bulduğu belge sayısını, "*a+b*" ilgili ya da ilgisiz erişilen toplam belge sayısını, "*a+c*" ise bir sorguya karşılık erişilen ya da erişilemeyen derlemdeki toplam ilgili belge sayısını verir. Anma, sistem tarafından erişilen ilgili belgelerin (a) derlemdeki toplam ilgili belgelere (a+c) oranını verir. Duyarlık ise sistem tarafından erişilen ilgili belgelerin (a) erişim çıktısında yer alan (ilgili ve ilgisiz) toplam belgelere (a+b) oranını verir. Bilgi erişim sistemlerinde anma ve duyarlık değerleri ne kadar yüksek olursa sistemin o kadar başarılı olduğu kabul edilir (Salton, 1989).

Tablo 1. İkili sınıflama tablosu

	İlgili	**İlgisiz**	
Erişilen	a	b	a+b
Erişilemeyen	c	d	c+d
	a+c	b+d	a+b+c+d

Bu çalışma kapsamında gerçekleştirilen test sonuçlarının değerlendirilmesinde, anma, duyarlık ölçüleri ile beraber anma ve duyarlığın harmonik ortalaması alınarak hesaplanan f-ölçü (f-measure[5]) değerinden de yararlanılmıştır.

[5] $f-\text{ölçü} = 2.\frac{anmaXduyarlık}{anma+duyarlık}$ formülü ile hesaplanır.

4. BÖLÜM

BULGULAR VE TARTIŞMA

4.1. HİKÂYE BAĞLANTI ALGILAMA BAŞARIM TESTLERİ

Daha önce Yöntem bölümünde açıklandığı gibi, hikâye bağlantı algılama görevinde başarım testleri gerçekleştirilirken vektör uzayı ve ilgi modeli yaklaşımları kullanılmıştır. Her bir yöntemin başarım testleri gerçekleştirilirken, öncelikle, eğitim belgeleri üzerinden uygun eşik değerler belirlenmiş ve sonraki aşamada bu eşik değerlere göre test belgeleri üzerinde başarım testleri gerçekleştirilmiştir.

Başarım testleri, haberleri temsil etmek için seçilen kelime sayılarına göre tekrarlanmıştır. Haberleri temsil etmek için seçilen kelime sayıları, 1 kelimeden başlamış ve 1000 kelimeye kadar kademeli olarak artırılarak devam etmiştir. Kelimeler seçilirken derlemin özellikleri dikkate alınarak *tf.idf* değeri en yüksek olan ilk *N* kelime kullanılmıştır. Başarım testlerinde uygulanan yöntemlerle ilgili olarak sonuçlar, ilgili terim sayılarına göre uygun eşik değerleri ile bu eşik değerlerindeki anma, duyarlık ve f-ölçü değerlerini göstermektedir. Sonuçlar eğitim ve test kümeleri için tablolar ve grafikler halinde sunulmaktadır.

Tablolarda "*TS*" haberleri temsil etmek için seçilen ve *tf.idf* değeri en yüksek olan terim (ya da kelime) sayısını, "*Eşik*" ilgili terim sayısında ilgililik değerlendirilmesinde kullanılan eşik değerini, "*Anma*", "*Duyarlık*" ve "*F-Ölçü*" ise seçilen eşik değerine göre elde edilen başarım ölçülerini göstermektedir. Eğitim kümesi üzerinde eşik değeri belirlenirken, anma ve duyarlığın birbirine en yakın olduğu nokta seçildiği için, tablolardan da görüleceği gibi, eğitim kümesi sonuçları için bu iki değer birbirine oldukça yakındır. Bununla birlikte, başarım testleri, eğitim aşamasında elde edilen bu eşik değerler temel alınarak gerçekleştirilmiştir.

4.1.1. Vektör Uzayı Modeli

Hikâye bağlantı algılama görevinin başarımı ile ilgili olarak uygulanan vektör uzayı yöntemi için elde edilen eğitim ve test sonuçları Tablo 2'de sunulmuştur.

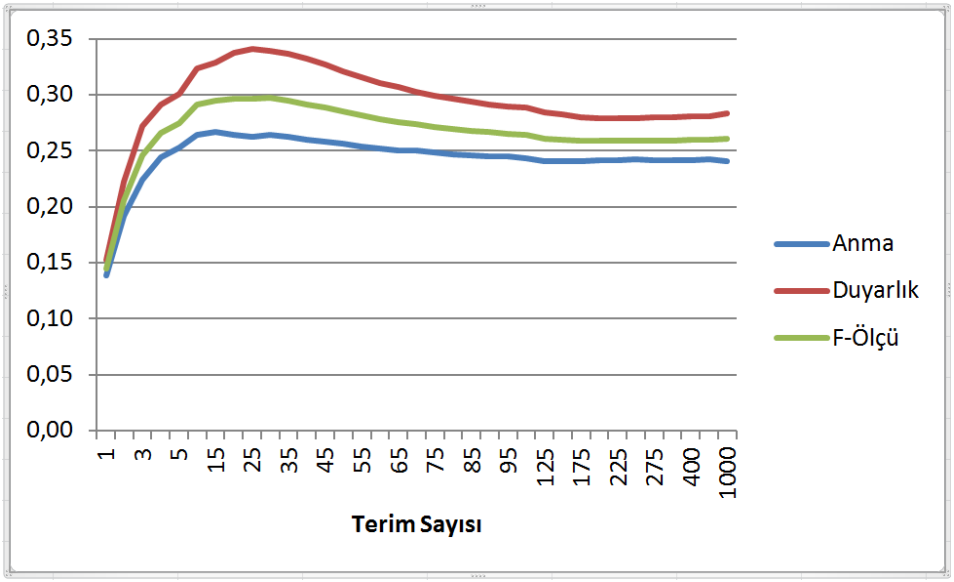
Vektör uzayı modeli için gerçekleştirilen testler sonucu, başarım metriği olarak f-ölçü değeri göz önüne alındığında, test derleminde en yüksek başarım 0,2970 olarak 30 terim için tespit edilmiştir. Bu f-ölçü değerinde anma 0,2642, duyarlık ise 0,3393 olarak gözlenmiştir. Diğer taraftan en yüksek anma değeri 0,2665 ile 15 terim, en yüksek duyarlık değeri ise 0,3406 ile 25 terim için elde edilmiştir. En yüksek anma ve duyarlık değerleri f-ölçü değerinin en yüksek olduğu noktadaki anma ve duyarlıkla karşılaştırıldığında önemli bir f-ölçü başarım farkı tespit edilememiştir (anma için %0,23 ve duyarlık için %0,13).

Bu sonuçlara ek olarak vektör uzayı modelinde belgeleri ifade etmek için kullanılan terim sayılarına bakıldığında, en yüksek başarımın 0,2970 f-ölçü değeri ile 30 terim için, en düşük başarımın ise 0,1451 f-ölçü değeri ile 1 terim için elde edildiği görülmektedir. Diğer taraftan bir belgeyi göstermek için 1000 terim (hemen hemen tüm terimler) kullanıldığı durumda f-ölçü değeri 0,2603 olarak gerçekleşmiştir. Bu değerlere göre en yüksek başarımın elde edildiği 30 terimde elde edilen f-ölçü değeri 1 terime göre %15,19, 1000 terime göre ise %3,67 daha yüksektir. Test sonuçlarındaki anma, duyarlık ve f-ölçü değerlerinin farklı terim sayılarına göre değişiminin gösterildiği Şekil 6 incelendiğinde, 10 ile 40 terim arasındaki değerlerde ciddi bir başarım farkı olmadığı ve bu noktalarda f-ölçü başarımının yataya yakın bir çizgi izlediği görülmektedir.

Tablo 2. Vektör uzayı modeli için eğitim ve test sonuçları[6]

	EĞİTİM				TEST		
TS	**Eşik**	**Anma**	**Duyarlık**	**F-Ölçü**	**Anma**	**Duyarlık**	**F-Ölçü**
1	0,5000	0,1535	0,1538	0,1536	0,1385	0,1523	0,1451
2	0,2330	0,2422	0,2429	0,2425	0,1922	0,2222	0,2061
3	0,1810	0,2911	0,2914	0,2912	0,2242	0,2720	0,2458
4	0,1410	0,3107	0,3104	0,3105	0,2442	0,2916	0,2658
5	0,1140	0,3239	0,3239	0,3239	0,2525	0,3006	0,2744
10	0,0600	0,3428	0,3425	0,3426	0,2644	0,3238	0,2911
15	0,0420	0,3503	0,3508	0,3505	**0,2665**	0,3286	0,2943
20	0,0340	0,3550	0,3547	0,3546	0,2641	0,3378	0,2964
25	0,0293	0,3550	0,3559	0,3554	0,2623	**0,3406**	0,2964
30	0,0262	**0,3556**	**0,3560**	**0,3558**	0,2642	0,3393	**0,2970**
35	0,0243	0,3546	0,3542	0,3544	0,2628	0,3365	0,2951
40	0,0231	0,3507	0,3519	0,3513	0,2595	0,3320	0,2913
45	0,0222	0,3473	0,3477	0,3475	0,2577	0,3267	0,2882
50	0,0215	0,3456	0,3443	0,3449	0,2564	0,3208	0,2850
55	0,0210	0,3430	0,3414	0,3422	0,2540	0,3153	0,2814
60	0,0207	0,3387	0,3398	0,3392	0,2516	0,3107	0,2781
65	0,0204	0,3352	0,3363	0,3357	0,2505	0,3067	0,2758
70	0,0201	0,3339	0,3327	0,3333	0,2500	0,3022	0,2737
75	0,0199	0,3317	0,3305	0,3311	0,2489	0,2989	0,2716
80	0,0198	0,3293	0,3293	0,3293	0,2471	0,2963	0,2694
85	0,0197	0,3282	0,3275	0,3278	0,2457	0,2937	0,2675
90	0,0196	0,3274	0,3262	0,3268	0,2453	0,2916	0,2664
95	0,0195	0,3256	0,3243	0,3249	0,2446	0,2893	0,2651
100	0,0195	0,3237	0,3237	0,3237	0,2437	0,2887	0,2643
125	0,0194	0,3186	0,3194	0,3190	0,2406	0,2846	0,2607
150	0,0193	0,3172	0,3169	0,3170	0,2406	0,2822	0,2597
175	0,0192	0,3166	0,3154	0,3160	0,2410	0,2795	0,2589
200	0,0192	0,3163	0,3157	0,3160	0,2412	0,2788	0,2587
225	0,0192	0,3166	0,3159	0,3162	0,2417	0,2787	0,2589
250	0,0192	0,3173	0,3161	0,3167	0,2420	0,2786	0,2590
275	0,0193	0,3165	0,3177	0,3171	0,2411	0,2799	0,2591
300	0,0193	0,3168	0,3178	0,3173	0,2414	0,2803	0,2594
400	0,0193	0,3180	0,3177	0,3177	0,2417	0,2806	0,2597
500	0,0193	0,3180	0,3176	0,3178	0,2420	0,2811	0,2601
1000	0,0194	0,3170	0,3187	0,3178	0,2409	0,2831	0,2603

[6] Verilen test sonuçları 11K030 numaralı proje sonuç raporunda 25 ve 26. sayfalarda sunulmuştur (Soydal ve Al, 2014).



Şekil 6. Vektör uzayı modeli test sonuçları başarım karşılaştırması

4.1.2. İlgi Modeli

Hikâye bağlantı algılama görevinin başarımı ile ilgili olarak uygulanan ilgi modeli yöntemi için elde edilen eğitim ve test sonuçları Tablo 3'te sunulmuştur.

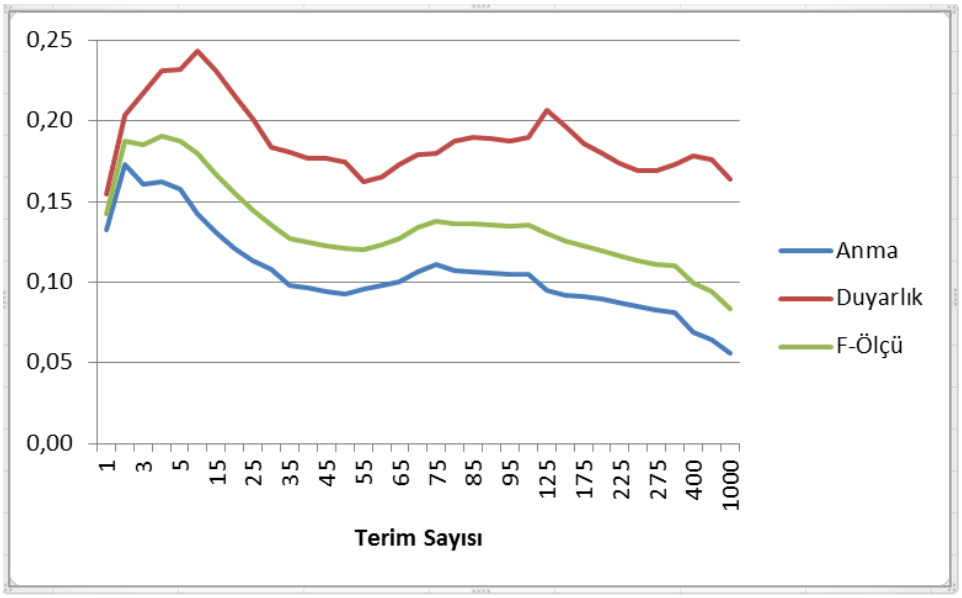
İlgi modeli için gerçekleştirilen testlerde en yüksek başarım 0,1910 f-ölçü değeri ile 4 terim için elde edilmiştir. Bu değer için anma 0,1625, duyarlık ise 0,2316 olarak gerçekleşmiştir. Buna ek olarak en yüksek anma değeri 0,1734 ile 2 terim için, en yüksek duyarlık değeri ise 0,2439 ile 10 terim için elde edilmiştir. Bu değerler f-ölçü değerinin en yüksek olduğu noktadaki anma ve duyarlık değerleri ile karşılaştırıldığında anmanın %1,09, duyarlığın ise %1,23 daha yüksek olduğu gözlenmiştir.

İlgi modelinde belgeleri ifade etmek için kullanılan terim sayılarına bakıldığında en yüksek başarımın 0,1910 f-ölçü değeri ile 4 terim için, en düşük başarımın ise 0,0839 ile 1000 terim için elde edildiği görülmektedir. İlgi modeli test sonuçlarındaki anma, duyarlık ve f-ölçü değerlerinin farklı terim sayılarına göre değişiminin gösterildiği Şekil 7 incelendiğinde, 2, 3, 4 ve 5 terim için ilgi modelinde en yüksek f-ölçü başarımına ulaşıldığı, ancak bu değerlerden sonra başarımın ciddi bir biçimde düşmeye başladığı görülmektedir.

Tablo 3. İlgi modeli için eğitim ve test sonuçları[7]

	EĞİTİM				TEST		
TS	**Eşik**	**Anma**	**Duyarlık**	**F-Ölçü**	**Anma**	**Duyarlık**	**F-Ölçü**
1	-5,9200	0,1567	0,1575	0,1571	0,1322	0,1544	0,1424
2	-6,6100	**0,2204**	**0,2200**	**0,2202**	**0,1734**	0,2039	0,1874
3	-6,7000	0,2162	0,2156	0,2159	0,1609	0,2179	0,1851
4	-6,7800	0,2140	0,2131	0,2135	0,1625	0,2316	**0,1910**
5	-6,8000	0,2186	0,2177	0,2181	0,1581	0,2318	0,1880
10	-6,7600	0,2064	0,2060	0,2062	0,1425	**0,2439**	0,1799
15	-6,6400	0,1886	0,1886	0,1886	0,131	0,2316	0,1673
20	-6,5000	0,1794	0,1781	0,1787	0,1213	0,2162	0,1554
25	-6,3000	0,1652	0,1679	0,1665	0,1131	0,2015	0,1449
30	-6,1300	0,1562	0,1588	0,1575	0,1079	0,1838	0,1360
35	-5,9200	0,1453	0,1453	0,1453	0,0979	0,1809	0,1270
40	-5,7800	0,1413	0,1415	0,1414	0,0965	0,1773	0,1249
45	-5,6600	0,1391	0,1402	0,1396	0,0941	0,1767	0,1228
50	-5,5600	0,1377	0,1389	0,1383	0,0930	0,1748	0,1214
55	-5,5000	0,1376	0,1355	0,1365	0,0960	0,1623	0,1206
60	-5,4400	0,1377	0,1380	0,1378	0,0980	0,1653	0,1231
65	-5,3900	0,1361	0,1383	0,1372	0,1005	0,1733	0,1272
70	-5,3600	0,1383	0,1406	0,1394	0,1069	0,1795	0,1340
75	-5,3300	0,1401	0,1396	0,1398	0,1115	0,1803	0,1378
80	-5,2800	0,1382	0,1394	0,1388	0,1073	0,1880	0,1366
85	-5,2500	0,1383	0,1387	0,1385	0,1062	0,1899	0,1362
90	-5,2200	0,1394	0,1363	0,1378	0,1057	0,189	0,1356
95	-5,1900	0,1389	0,1405	0,1397	0,1049	0,1878	0,1346
100	-5,1700	0,1405	0,1399	0,1402	0,1051	0,1898	0,1353
125	-5,0400	0,1413	0,1428	0,1420	0,0949	0,2068	0,1301
150	-4,9500	0,1421	0,1404	0,1412	0,0923	0,1970	0,1257
175	-4,8800	0,1409	0,1416	0,1412	0,0915	0,1864	0,1227
200	-4,8100	0,1389	0,1414	0,1401	0,0896	0,1797	0,1196
225	-4,7500	0,1372	0,1362	0,1367	0,0875	0,1737	0,1164
250	-4,7000	0,1341	0,1387	0,1364	0,0852	0,1691	0,1133
275	-4,6600	0,1319	0,1351	0,1335	0,0830	0,1690	0,1113
300	-4,6200	0,1306	0,1296	0,1301	0,0810	0,1732	0,1103
400	-4,5000	0,1181	0,1197	0,1189	0,0691	0,1783	0,0996
500	-4,4400	0,1111	0,1125	0,1118	0,0646	0,1760	0,0945
1000	-4,3800	0,1024	0,1032	0,1028	0,0563	0,1643	0,0839

[7] Verilen test sonuçları 11K030 numaralı proje sonuç raporunda 27 ve 28. sayfalarda sunulmuştur (Soydal ve Al, 2014).



Şekil 7. İlgi modeli test sonuçları başarım karşılaştırması

4.1.3. Birleştirilmiş Sonuçlar

Hikâye bağlantı algılama görevinin başarımı ile ilgili olarak uygulanan vektör uzayı ve ilgi modeli yöntemlerinden elde edilen sonuçların AND ve OR mantıksal operatörleri ile birleştirildikten sonra elde edilen başarım değerleri Tablo 4'te sunulmuştur.

Yöntemlerin AND birleşimlerinde, vektör uzayı ve ilgi modeli erişim çıktılarında, haber eşleşmeleri için her iki yöntemin de ilgili dediği çıktılar ilgili olarak kabul edilmiş, bunun dışındaki durumlarda, çıktı, ilgisiz olarak işaretlenmiştir. OR birleşimlerinde ise, yöntemlerden bir tanesi bile eşleşmeyi ilgili olarak belirlediyse, sonuç, ilgili olarak kabul edilmiş, her iki yöntemin de ilgisiz kabul ettiği çıktılar, ilgisiz olarak işaretlenmiştir. Yöntemlerin sonuçlarının birleştirildiği başarım değerlerine bakıldığında, beklenen sonuçların görüldüğü söylenebilir. Bu kapsamda AND birleşimlerinde en yüksek başarımın 0,2216'lık f-ölçü değeri ile (anma 0,1507 ve duyarlık 0,4183) 4 terim için sağlandığı görülmektedir. AND birleşimi vektör uzayı modeli ile en yüksek f-ölçü başarımının elde edildiği 30 terimde duyarlık değerini %8,62, ilgi modeli ile en yüksek f-ölçü başarımının elde edildiği 4 terimde ise duyarlık değerini %18,67 oranında artırmıştır. Buna karşılık AND birleşiminde anma değerleri 30 terim vektör uzayı modeli için %16,92 ve 4 terim ilgi modeli için %1,18 düşmüştür. OR birleşiminde ise en yüksek başarım 0,2641'lik f-ölçü değeri ile (anma 0,2762 ve

duyarlık 0,2531) 15 terim için elde edilmiştir. Buna göre sonuçları OR ile birleştirme, vektör uzayı modeli ile en yüksek başarımın elde edildiği 30 terimde anma değerini %1,3, ilgi modeli ile en yüksek başarımın elde edildiği 4 terimde ise anma değerini %9,35 oranında artırmıştır. Diğer taraftan OR birleşiminde duyarlık değerleri 30 terim vektör uzayı modeli için %9,7 ve 4 terim ilgi modeli için %1,44 düşmüştür.

4.1.4. Uygulanan Yöntemlerin Karşılaştırılması

Hikâye bağlantı algılama görevinin başarımının belirlenmesinde kullanılan vektör uzayı modeli ve ilgi modeli ile bunların AND ve OR mantıksal operatörleri kullanılarak birleştirilmiş biçimlerinden elde edilen başarım ölçütleri, yöntemleri daha rahat karşılaştırabilmek için, Tablo 5, 6 ve 7'de sunulmaktadır. Tablolarda "TS" terim sayısını, "F-Ölçü (VUM)" vektör uzayı modeli için f-ölçü değerlerini, "F-Ölçü (İM)" ilgi modeli için f-ölçü değerlerini, "F-Ölçü (AND)" bu iki yöntemin AND birleşiminden, "F-Ölçü (OR)" ise bu iki yöntemin OR birleşiminden elde edilen f-ölçü değerlerini, "Anma (VUM)" vektör uzayı modeli için anma değerlerini, "Anma (İM)" ilgi modeli için anma değerlerini, "Anma (AND)" bu iki yöntemin AND birleşiminden, "Anma (OR)" ise bu iki yöntemin OR birleşiminden elde edilen anma değerlerini, "Duyarlık (VUM)" vektör uzayı modeli için duyarlık değerlerini, "Duyarlık (İM)" ilgi modeli için duyarlık değerlerini, "Duyarlık (AND)" bu iki yöntemin AND birleşiminden, "Duyarlık (OR)" ise bu iki yöntemin OR birleşiminden elde edilen duyarlık değerlerini ifade etmektedir. MAKS ve MİN ise ilgili yöntemler için erişilen en yüksek ve en düşük değerleri ifade etmektedir.

Tablo 5 incelendiğinde, hikâye bağlantı algılama görevinde f-ölçü başarım değerlerine göre en yüksek başarımlar; vektör uzayı modelinde 0,2970 ile 30 terim için, ilgi modelinde 0,1910 ile 4 terim için, AND birleşimlerinde 0,2216 ile 4 terim için ve OR birleşimlerinde 0,2641 ile 15 terim için elde edilmiştir.

Hikâye bağlantı algılama görevinde, vektör uzayı modeli ile elde edilen f-ölçü başarımının ne ilgi modelinde ne de yöntemlerin AND, OR birleşimlerinde yakalanamadığı görülmektedir.

Tablo 4. Vektör uzayı ve ilgi modeli için AND ve OR birleşim sonuçları[8]

	AND			OR		
TS	**Anma**	**Duyarlık**	**F-Ölçü**	**Anma**	**Duyarlık**	**F-Ölçü**
1	0,0807	0,2680	0,1240	0,1900	0,1297	0,1542
2	**0,1587**	0,3475	0,2179	0,2068	0,1643	0,1831
3	0,1474	0,3886	0,2137	0,2377	0,2009	0,2178
4	0,1507	0,4183	**0,2216**	0,2560	0,2172	0,2350
5	0,1469	0,4180	0,2174	0,2636	0,2252	0,2429
10	0,1340	0,4259	0,2038	0,2729	0,2513	0,2617
15	0,1213	0,4251	0,1887	0,2762	**0,2531**	**0,2641**
20	0,1103	0,4307	0,1757	0,2750	0,2531	0,2636
25	0,1016	0,4282	0,1642	0,2738	0,2502	0,2615
30	0,0949	0,4255	0,1553	**0,2771**	0,2425	0,2587
35	0,0860	0,4204	0,1427	0,2747	0,2459	0,2595
40	0,0841	0,4126	0,1397	0,2719	0,2423	0,2563
45	0,0825	0,4065	0,1371	0,2694	0,2408	0,2543
50	0,0818	0,4075	0,1363	0,2675	0,2367	0,2512
55	0,0827	0,4069	0,1374	0,2674	0,2240	0,2438
60	0,0838	0,4212	0,1397	0,2659	0,2208	0,2413
65	0,0855	0,4357	0,1429	0,2656	0,2212	0,2413
70	0,0900	0,4452	0,1497	0,2669	0,2187	0,2404
75	0,0926	0,4597	0,1541	0,2677	0,2143	0,2381
80	0,0899	0,4776	0,1513	0,2645	0,2174	0,2386
85	0,0894	0,4883	0,1512	0,2624	0,2165	0,2372
90	0,0888	0,4966	0,1506	0,2622	0,2146	0,2360
95	0,0875	0,5052	0,1491	0,2621	0,2129	0,2349
100	0,0873	0,5131	0,1493	0,2615	0,2130	0,2348
125	0,0793	0,5642	0,1390	0,2562	0,2201	0,2368
150	0,0745	0,6113	0,1329	0,2583	0,2155	0,2349
175	0,0704	0,6380	0,1269	0,2621	0,2109	0,2337
200	0,0666	0,6592	0,1209	0,2643	0,2093	0,2336
225	0,0632	0,6814	0,1157	0,2660	0,2081	0,2335
250	0,0601	0,7011	0,1107	0,2671	0,2076	0,2336
275	0,0574	0,7154	0,1063	0,2667	0,2096	0,2347
300	0,0549	0,7256	0,1020	0,2675	0,2135	0,2374
400	0,0467	0,7618	0,0880	0,2641	0,2224	0,2415
500	0,0432	0,7795	0,0819	0,2634	0,2246	0,2425
1000	0,0393	**0,7987**	0,0749	0,2579	0,2254	0,2405
MAKS	**0,1587**	**0,7987**	**0,2216**	**0,2771**	**0,2531**	**0,2641**
MİN	**0,0393**	**0,2680**	**0,0749**	**0,1900**	**0,1297**	**0,1542**

[8] Verilen test sonuçları 11K030 numaralı proje sonuç raporunda 29 ve 30. sayfalarda sunulmuştur (Soydal ve Al, 2014).

Tablo 5. Yöntemlerin f-ölçü değeri karşılaştırmaları

TS	F-Ölçü (VUM)	F-Ölçü (İM)	F-Ölçü (AND)	F-Ölçü (OR)
1	0,1451	0,1424	0,1240	0,1542
2	0,2061	0,1874	0,2179	0,1831
3	0,2458	0,1851	0,2137	0,2178
4	0,2658	**0,1910**	**0,2216**	0,2350
5	0,2744	0,1880	0,2174	0,2429
10	0,2911	0,1799	0,2038	0,2617
15	0,2943	0,1673	0,1887	**0,2641**
20	0,2964	0,1554	0,1757	0,2636
25	0,2964	0,1449	0,1642	0,2615
30	**0,2970**	0,1360	0,1553	0,2587
35	0,2951	0,1270	0,1427	0,2595
40	0,2913	0,1249	0,1397	0,2563
45	0,2882	0,1228	0,1371	0,2543
50	0,2850	0,1214	0,1363	0,2512
55	0,2814	0,1206	0,1374	0,2438
60	0,2781	0,1231	0,1397	0,2413
65	0,2758	0,1272	0,1429	0,2413
70	0,2737	0,1340	0,1497	0,2404
75	0,2716	0,1378	0,1541	0,2381
80	0,2694	0,1366	0,1513	0,2386
85	0,2675	0,1362	0,1512	0,2372
90	0,2664	0,1356	0,1506	0,2360
95	0,2651	0,1346	0,1491	0,2349
100	0,2643	0,1353	0,1493	0,2348
125	0,2607	0,1301	0,1390	0,2368
150	0,2597	0,1257	0,1329	0,2349
175	0,2589	0,1227	0,1269	0,2337
200	0,2587	0,1196	0,1209	0,2336
225	0,2589	0,1164	0,1157	0,2335
250	0,2590	0,1133	0,1107	0,2336
275	0,2591	0,1113	0,1063	0,2347
300	0,2594	0,1103	0,1020	0,2374
400	0,2597	0,0996	0,0880	0,2415
500	0,2601	0,0945	0,0819	0,2425
1000	0,2603	0,0839	0,0749	0,2405
MAKS	**0,2970**	**0,1910**	**0,2216**	**0,2641**
MİN	**0,1451**	**0,0839**	**0,0749**	**0,1542**

Yöntemlerin AND birleşimlerinin 2 terim kullanılan durum dışında, vektör uzayı modelinin f-ölçü başarımını genelde düşürdüğü, ilgi modeli sonuçlarında ise belirli terim sayılarında sınırlı bir iyileşme sağladığı görülmektedir. AND birleşimi vektör uzayı modeli ile en yüksek f-ölçü başarımının elde edildiği 30 terimde başarımı %14,18 düşürürken, ilgi modelinde en yüksek başarımın elde edildiği 4 terimde başarımı %3,6 artırmıştır.

Yöntemlerin OR birleşimlerinin kullanıldığı sonuçlarda ise, vektör uzayı modelinde 1 terimin kullanıldığı durum dışında, bu yöntemin başarımında düşüş olmuştur. Diğer taraftan, OR birleşimi ilgi modelinde, hemen hemen tüm senaryolarda başarımı ciddi oranda artırmıştır. OR birleşimi, vektör uzayı modeli ile en yüksek f-ölçü başarımının elde edildiği 30 terimde başarımı %3,84 düşürürken, ilgi modelinde en yüksek f-ölçü başarımının elde edildiği 4 terimde başarımı %4,4 artırmıştır.

Tablo 6'da hikâye bağlantı algılama görevinde uygulanan yöntemler ve bunların AND, OR mantıksal operatörleri kullanılarak birleştirilmiş biçimlerinden elde edilen anma değerleri sunulmuştur.

Yöntemlerin anma değerleri incelendiğinde, hikâye bağlantı algılama görevinde anma başarım değerlerine göre en yüksek başarımlar; vektör uzayı modelinde 0,2665 değeri ile 15 terim için, ilgi modelinde 0,1734 değeri ile 2 terim için, AND birleşimlerinde 0,1587 değeri ile 2 terim için ve OR birleşimlerinde 0,2771 değeri ile 30 terim için elde edildiği görülmektedir.

Yöntemlerin AND ve OR birleşimleri incelendiğinde, beklendiği gibi, AND birleşiminin hem vektör uzayı hem de ilgi modelinde tüm terim sayılarında anma değerini düşürdüğü, OR birleşiminin ise her iki yöntemde de anma değerlerini artırdığı görülmüştür.

Tablo 6. Yöntemlerin anma değeri karşılaştırmaları

TS	Anma (VUM)	Anma (İM)	Anma (AND)	Anma(OR)
1	0,1385	0,1322	0,0807	0,1900
2	0,1922	**0,1734**	**0,1587**	0,2068
3	0,2242	0,1609	0,1474	0,2377
4	0,2442	0,1625	0,1507	0,2560
5	0,2525	0,1581	0,1469	0,2636
10	0,2644	0,1425	0,1340	0,2729
15	**0,2665**	0,1310	0,1213	0,2762
20	0,2641	0,1213	0,1103	0,2750
25	0,2623	0,1131	0,1016	0,2738
30	0,2642	0,1079	0,0949	**0,2771**
35	0,2628	0,0979	0,0860	0,2747
40	0,2595	0,0965	0,0841	0,2719
45	0,2577	0,0941	0,0825	0,2694
50	0,2564	0,0930	0,0818	0,2675
55	0,2540	0,0960	0,0827	0,2674
60	0,2516	0,0980	0,0838	0,2659
65	0,2505	0,1005	0,0855	0,2656
70	0,2500	0,1069	0,0900	0,2669
75	0,2489	0,1115	0,0926	0,2677
80	0,2471	0,1073	0,0899	0,2645
85	0,2457	0,1062	0,0894	0,2624
90	0,2453	0,1057	0,0888	0,2622
95	0,2446	0,1049	0,0875	0,2621
100	0,2437	0,1051	0,0873	0,2615
125	0,2406	0,0949	0,0793	0,2562
150	0,2406	0,0923	0,0745	0,2583
175	0,2410	0,0915	0,0704	0,2621
200	0,2412	0,0896	0,0666	0,2643
225	0,2417	0,0875	0,0632	0,2660
250	0,2420	0,0852	0,0601	0,2671
275	0,2411	0,0830	0,0574	0,2667
300	0,2414	0,0810	0,0549	0,2675
400	0,2417	0,0691	0,0467	0,2641
500	0,2420	0,0646	0,0432	0,2634
1000	0,2409	0,0563	0,0393	0,2579
MAKS	**0,2665**	**0,1734**	**0,1587**	**0,2771**
MİN	**0,1385**	**0,0563**	**0,0393**	**0,1900**

AND birleşiminde vektör uzayı modelinde anma değerinin en yüksek olduğu 15 terimde %14,52, ilgi modelinde anma değerinin en yüksek olduğu 2 terimde %1,47'lik düşüş olmuştur. AND birleşiminde haber çiftlerinin aynı konuda olup olmadıklarını belirleyebilmek için her iki yöntemin de "ilgili" kararı vermesi gerekmektedir. Bu nedenle, AND birleşimlerinde anma değerinin bağımsız uygulanan yöntemlerden daha düşük çıkması, beklenen bir durumdur.

Diğer taraftan OR birleşiminde vektör uzayı modelinde anma değerinin en yüksek olduğu 15 terimde %0,97, ilgi modelinde anma değerinin en yüksek olduğu 2 terimde %3,35'lik artış olmuştur. OR birleşiminde haber çiftlerinin aynı konuda olup olmadıklarını belirleyebilmek için uygulanan yöntemlerden birisinin "ilgili" kararı vermesi yeterlidir. Bu nedenle OR birleşimlerinde, anma değerinin, bağımsız uygulanan yöntemlerden daha yüksek çıkması beklenen bir durumdur.

Tablo 7'de hikâye bağlantı algılama görevinde uygulanan yöntemler ve bunların AND ve OR mantıksal operatörleri kullanılarak birleştirilmiş biçimlerinden elde edilen duyarlık değerleri sunulmuştur.

Yöntemlerin duyarlık değerleri incelendiğinde, hikâye bağlantı algılama görevinde duyarlık başarım değerlerine göre en yüksek başarımlar; vektör uzayı modelinde 0,3406 değeri ile 25 terim için, ilgi modelinde 0,2439 değeri ile 10 terim için, AND birleşimlerinde 0,7987 değeri ile 1000 terim için ve OR birleşimlerinde 0,2531 değeri ile 15 terim için elde edildiği görülmektedir.

Yöntemlerin AND ve OR birleşimleri incelendiğinde, AND birleşiminin hem vektör uzayı hem de ilgi modelinde, tüm terim sayılarında duyarlık değerini yükselttiğini, OR birleşiminin ise, vektör uzayı yönteminde duyarlığın düşmesine neden olurken, ilgi modelinde özellikle 5 terimden sonrası için sınırlı da olsa duyarlık artışına neden olduğu görülmektedir.

Tablo 7. Yöntemlerin duyarlık değeri karşılaştırmaları

TS	Duyarlık (VUM)	Duyarlık (İM)	Duyarlık (AND)	Duyarlık(OR)
1	0,1523	0,1544	0,2680	0,1297
2	0,2222	0,2039	0,3475	0,1643
3	0,2720	0,2179	0,3886	0,2009
4	0,2916	0,2316	0,4183	0,2172
5	0,3006	0,2318	0,4180	0,2252
10	0,3238	**0,2439**	0,4259	0,2513
15	0,3286	0,2316	0,4251	**0,2531**
20	0,3378	0,2162	0,4307	0,2531
25	**0,3406**	0,2015	0,4282	0,2502
30	0,3393	0,1838	0,4255	0,2425
35	0,3365	0,1809	0,4204	0,2459
40	0,3320	0,1773	0,4126	0,2423
45	0,3267	0,1767	0,4065	0,2408
50	0,3208	0,1748	0,4075	0,2367
55	0,3153	0,1623	0,4069	0,2240
60	0,3107	0,1653	0,4212	0,2208
65	0,3067	0,1733	0,4357	0,2212
70	0,3022	0,1795	0,4452	0,2187
75	0,2989	0,1803	0,4597	0,2143
80	0,2963	0,1880	0,4776	0,2174
85	0,2937	0,1899	0,4883	0,2165
90	0,2916	0,1890	0,4966	0,2146
95	0,2893	0,1878	0,5052	0,2129
100	0,2887	0,1898	0,5131	0,2130
125	0,2846	0,2068	0,5642	0,2201
150	0,2822	0,1970	0,6113	0,2155
175	0,2795	0,1864	0,6380	0,2109
200	0,2788	0,1797	0,6592	0,2093
225	0,2787	0,1737	0,6814	0,2081
250	0,2786	0,1691	0,7011	0,2076
275	0,2799	0,1690	0,7154	0,2096
300	0,2803	0,1732	0,7256	0,2135
400	0,2806	0,1783	0,7618	0,2224
500	0,2811	0,1760	0,7795	0,2246
1000	0,2831	0,1643	**0,7987**	0,2254
MAKS	**0,3406**	**0,2439**	**0,7987**	**0,2531**
MİN	**0,1523**	**0,1544**	**0,2680**	**0,1297**

AND birleşiminde, vektör uzayı modelinde, duyarlık değerinin en yüksek olduğu 25 terimde %8,76, ilgi modelinde duyarlık değerinin en yüksek olduğu 10 terimde %18,2'lik artış olmuştur. AND birleşiminde, haber çiftlerinin aynı konuda olduklarını belirleyebilmek için, her iki yöntemin de "ilgili" kararı vermesi gerekmektedir. Bu nedenle AND birleşimlerinin daha seçici olduğunu, doğal olarak da, duyarlık değerlerinin bağımsız uygulanan yöntemlerden daha yüksek çıkmasının normal olduğunu söylemek mümkündür.

Diğer taraftan OR birleşiminde, vektör uzayı modelinde, duyarlık değerinin en yüksek olduğu 25 terimde %9,04 azalış, ilgi modelinde, duyarlık değerinin en yüksek olduğu 10 terimde %0,73'lük artış olmuştur. OR birleşimlerinde haber çiftlerinin aynı konuda olup olmadıklarını belirleyebilmek için uygulanan yöntemlerden birisinin "ilgili" kararı vermesi yeterlidir. Bu nedenle, OR birleşimlerinde erişim çıktısında hatalı eşleşmelerle karşılaşma olasılığı her zaman daha yüksek olarak görülür.

4.2. KONU İZLEME BAŞARIM TESTLERİ

Daha önce Yöntem bölümünde açıklandığı gibi, konu izleme görevinde başarım testleri gerçekleştirilirken, k-ortalamalar, vektör uzayı ve ilgi modeli yöntemleri kullanılmıştır. Her bir yöntemin başarım testleri gerçekleştirilirken, öncelikle eğitim belgeleri üzerinden uygun eşik değerler belirlenmiş ve sonraki aşamada, bu eşik değerlere göre test belgeleri üzerinde başarım testleri gerçekleştirilmiştir. Başarım testleri gerçekleştirilirken, haberleri ifade etmek için seçilen terim sayıları, daha güçlü konu modelleri yaratmak amacıyla, sınırlandırılmamış olup tüm terimler kullanılmıştır.

Ayrıca bu görevle ilgili olarak, uygun eşik değeri belirleme yönteminin belirlenmesi için bir dizi test uygulanmış olup, bu testlerin sonuçları da bu bölüm içerisinde tablolar halinde sunulmaktadır. Tablolarda "*Eşik*" ilgililik değerlendirmesinde kullanılan eşik değerini, "*Anma*", "*Duyarlık*" ve "*F-Ölçü*" ise seçilen eşik değerine göre elde edilen başarım ölçülerini göstermektedir.

4.2.1. Uygun Eşik Değeri Belirleme Yöntemi

Konu izleme görevinde, haberlerin daha önce belirlenmiş olan konularla ilgili olup olmadıkları saptanırken, bir eşik değerinden yararlanılmaktadır. Başlangıçta yaratılan küme merkezi vektörleri, haber vektörleri ile karşılaştırılmakta ve belirlenen eşik değerinin üzerinde benzerlik skoruna sahip olan haberler, konuyla ilgili olarak kabul edilmektedir. Hikâye bağlantı algılama görevinde kullanılan yöntemlerde de benzer bir yaklaşım olmasına rağmen, konu izlemede kullanılan yöntemlerde, konu kümeleri ve küme merkezi vektörleri kullanıldığı için farklı kıstaslara göre eşik değeri seçmek mümkündür. Bu kapsamda, eğitim belgelerine göre yaratılan kümelerde, küme merkezi vektörüne en yakın belgenin uzaklığı, en uzak belgenin uzaklığı ya da kümeye ait tüm belgeşerin ortalama uzaklığı gibi farklı yöntemler kullanarak da eşik değeri seçmek mümkündür.

Bu kapsamda en uygun eşik seçme yöntemini belirlemek amacıyla bir dizi test yapılmıştır. Testler, BilCol-2005 derleminde 80 konu başlığında olan 5883 haberin 1961 tanesi eğitim, 3922 tanesi test belgesi olarak kabul edilerek gerçekleştirilmiştir. Eğitim kümesinde, her bir konu başlığı ile ilgili olarak, tarih sırasına göre ilk 4 belge kullanılarak küme merkezi vektörleri oluşturulmuş, daha sonra 3922 haber kullanılarak, farklı senaryoların başarımları test edilmiştir.

Gerçekleştirilen başarım testleri sonucunda elde edilen değerler, Tablo 8'de sunulmuştur. Tablo üzerinde harflerle kodlanmış olan yöntemler şu şekildedir:

A: Eşik değeri, eğitim belgelerinde anma/duyarlık değerlerinin en yüksek olduğu nokta olarak seçilmiştir.

B: Eşik değeri, tüm konular için küme merkezi vektörlerine en uzak olan eğitim belgesinin mesafesi alınarak seçilmiştir.

C: Eşik değeri, her bir konu merkezi vektörüne, o konuyla ilgili eğitim belgelerinin mesafelerinin ortalamaları alınarak seçilmiştir.

D: Eşik değeri, her bir konu için ayrı ayrı hesaplanmıştır. Her bir konunun küme merkezi vektörüne en uzak olan o konuyla ilgili belgenin mesafesi, eşik olarak seçilmiştir.

E: Küme merkezi vektörlerine en yakın belgenin mesafesi, eşik olarak seçilmiştir.

Anma ve duyarlığın en yüksek olduğu noktaya göre eşik değer belirlenmesi yönteminden elde edilen f-ölçü başarımı, diğer yöntemlerden ciddi anlamda yüksektir (Tablo 8). Bu kapsamda, konu izleme ile ilgili olarak gerçekleştirilen tüm testlerde eşik değeri belirleme yöntemi olarak, eğitim kümesinde anma ve duyarlığın en yüksek olduğu noktaya göre eşik değeri belirleme yöntemi kullanılmıştır.

Tablo 8. Konu izleme görevi için eşik değer belirleme yöntemi sonuçları[9]

	A	B	C	D	E
Eşik Değer	0,8660	0,9975	0,7043	Dinamik[10]	0,4308
Anma	0,6686	1,0000	0,1355	0,8206	0,0090
Duyarlık	0,9012	0,0130	0,9963	0,2650	1,0000
F-Ölçü	0,7677	0,0257	0,2386	0,4006	0,0178

4.2.2. Uygulanan Yöntemlerin Test Sonuçları

Konu izleme görevinde erişim fonksiyonu olarak kullanılan k-ortalamalar, vektör uzayı ve ilgi modellerinin başarım test sonuçları Tablo 9'da sunulmuştur. Uygulanan yöntemlerle ilgili olarak eşik, anma, duyarlık ve f-ölçü değerleri eğitim ve test belgeleri için ayrı ayrı gösterilmiştir.

K-ortalamalar yönteminde haber içeriklerini göstermek için tüm terimlerin kullanıldığı durumda 0,7677'lik f-ölçü değerine ulaşıldığı görülmektedir. Bu f-ölçü değeri için 0,6686'lık bir anma ve 0,9012 gibi oldukça yüksek bir duyarlık değeri elde edilmiştir.

[9] Verilen test sonuçları 11K030 numaralı proje sonuç raporunda 36. sayfada sunulmuştur (Soydal ve Al, 2014).

[10] Bu yöntemde eşik değeri her bir konu için ayrı ayrı belirlenir. Konu modelleri değiştikçe eşik değeri de dinamik olarak tekrar hesaplanmaktadır.

Tablo 9. Konu izleme görevi için uygulanan test sonuçları[11]

	K-ortalamalar		Vektör Uzayı Modeli		İlgi Modeli	
	Eğitim	**Test**	**Eğitim**	**Test**	**Eğitim**	**Test**
Eşik Değer	0,8660	0,8660	0,8754	0,8754	-6,07	-6,07
Anma	0,7929	0,6686	0,7628	0,6341	0,5843	0,5060
Duyarlık	0,7941	0,9012	0,7632	0,8698	0,5827	0,5124
F-Ölçü	0,7935	0,7677	0,7630	0,7335	0,5835	0,5092

Konu izleme görevi için vektör uzayı modelinin uygulandığı yöntemde, haber içeriklerini göstermek için tüm terimlerin kullanıldığı durumda 0,7335'lik f-ölçü değerine ulaşıldığı görülmektedir. Bu f-ölçü değeri için 0,6341'lık bir anma ve 0,8698'lik duyarlık değeri elde edilmiştir.

İlgi modelinin uygulandığı yöntemde, haber içeriklerini göstermek için tüm terimlerin kullanıldığı durumda 0,5092'lik f-ölçü değerine ulaşıldığı görülmektedir. Bu f-ölçü değeri için 0,5060'lık bir anma ve 0,5124'lük duyarlık değeri elde edilmiştir.

Yöntemlerin konu izleme görevindeki başarımlarına bakıldığında, konu izleme görevi için f-ölçü değerlerine göre en başarılı yöntemin, k-ortalamalar yöntemi olduğu görülmektedir. K-ortalamalar yöntemi, vektör uzayı modelinden %3,42, ilgi modelinden ise %25,85 oranında daha yüksek f-ölçü değeri üretmiştir.

Yöntemlerin anma değerlerine bakıldığında; k-ortalamalar yönteminin, vektör uzayı modelinden %3,45, ilgi modelinden %16,26 oranında daha başarlı sonuçlar ürettiği görülmüştür. Benzer başarım farkları, duyarlık değerleri için de gözlenmiş olup k-ortalamalar yöntemi, vektör uzayı modelinden %3,14, ilgi modelinden ise %38,88 oranında daha yüksek duyarlık değerleri üretmiştir.

4.3. BULGULARIN DEĞERLENDİRİLMESİ

Bu çalışma kapsamında elde edilen bulgular, TDT programında, Türkçe derlemler üzerinde gerçekleştirilen çalışmaların sınırlı olması açısından son derece önemlidir.

[11] Verilen test sonuçları 11K030 numaralı proje sonuç raporunda 36 ve 37. sayfalarda sunulmuştur (Soydal ve Al, 2014).

Gerçekleştirilen bu çalışma, özellikle hikâye bağlantı algılama görevi ile ilgili olarak Türkçe bir derlem üzerinde gerçekleştirilmiş olan bilinen ilk çalışmadır. Bu bağlamda, hikâye bağlantı algılama görevi için elde edilen bulgular, gerek belge gösterimi, gerekse erişim fonksiyonu açısından İngilizce ya da çok dilli olarak oluşturulmuş olan TDT derlemi üzerinde gerçekleştirilmiş çalışmalarla karşılaştırılarak değerlendirilmiştir.

2004 yılında gerçekleştirilen ayrıntılı bir çalışmada; araştırmacılar, TDT görevleri için yaygın olarak kullanılan yöntemlerin, çok dilli olarak oluşturulmuş olan TDT-4 ve TDT-5 derlemleri üzerindeki başarımlarını ortaya koymuşlardır (Connell ve diğerleri, 2004). Araştırmacılar tarafından raporlanan sonuçlara göre; hikâye bağlantı algılama görevinde, farklı dillerdeki haberler İngilizceye çevrildikten sonra yapılan başarım testlerinde, vektör uzayı modeli, ilgi modeline göre ortalama %0,65 daha başarılı sonuçlar üretmiştir. Diğer taraftan, farklı dillerdeki haberler İngilizceye çevrilmeden, yöntemler doğal dildeki haberler üzerinde uygulandığında, ilgi modeli, vektör uzayı modeline göre %0,48 daha başarılı sonuçlar üretmiştir. Bu çalışmadan elde edilen sonuçlar, vektör uzayı ve ilgi modeli arasındaki başarım farkının çok yüksek olmadığını, hatta bu farkın göz ardı edilebileceğini göstermektedir. Diğer taraftan Türkçe BilCol-2005 derlemi üzerinde gerçekleştirdiğimiz başarım testleri, f-ölçü değerlerine göre karşılaştırıldığında, vektör uzayı modelinin, ilgi modeline göre ciddi oranda daha başarılı sonuçlar ürettiği görülmüştür.

Aynı çalışmada (Connell ve diğerleri, 2004), konu izleme görevi ile ilgili olarak başarım testleri, eğiticili ve eğiticisiz olmak üzere iki senaryo üzerinde gerçekleştirilmiştir. Eğiticili senaryoda, küme merkezi vektörlerini oluşturmak için başlangıçta belirli sayıda haber kullanılırken, eğiticisiz senaryoda küme merkezi vektörleri tek haber kullanılarak oluşturulmuştur. Araştırmacıların elde ettiği sonuçlar, ilgi modelinin, vektör uzayı modeline göre, eğiticisiz senaryoda %0,57, eğiticili senaryoda ise %14,10 oranında daha başarılı olduğunu göstermektedir. Buna ek olarak en başarılı eğiticili yöntem ile en başarılı eğiticisiz yöntem arasında %8,75 oranında başarım farkı rapor edilmiştir. Bu sonuç, konu izleme görevinde güçlü konu modelleri oluşturmanın ne kadar önemli olduğunu açık bir biçimde göstermektedir. Konu izleme görevi ile ilgili olarak, BilCol-2005 derlemi üzerinde gerçekleştirdiğimiz testler ise, en başarılı yöntem olarak k-ortalamalar yöntemini işaret etmektedir. Hatta bu yöntemden

elde edilen başarım, ilgi modelinin oldukça üzerinde çıkmıştır. Buna ek olarak, konu izleme görevi ile ilgili olarak Türkçe bir derlem üzerinde başarım testlerinin ilk olarak gerçekleştirildiği çalışmada (Can ve diğerleri, 2010), araştırmacılar dil ya da ilgi modelleri yöntemlerini uygulamamışlar, ancak uygulanan yöntemler arasında vektör uzayı modelinden en yüksek başarımı elde etmişlerdir.

Belge gösterimi tarafında ise, TDT derlemi üzerinde gerçekleştirilen akademik çalışmalar, haberleri ifade etmek için 30 terim ya da üzerinde bir sayının yeterli olacağını göstermektedir (Lavrenko ve diğerleri, 2002). Buna göre hikâye bağlantı algılama görevinde, bir haberi ifade etmek için *tf.idf* değeri en yüksek olan ilk 30 terim seçilerek en başarılı sonuçlara ulaşılırken, 30 terimden az sayıda terim seçilmesi halinde başarımın düştüğü, 30 terimden fazla terim seçilmesi halinde ise başarımda anlamlı bir artış olmadığı rapor edilmiştir (Lavrenko ve diğerleri, 2002). Diğer taraftan BilCol-2005 derlemi üzerinde gerçekleştirilen testler, vektör uzayı modelinde 10-40, ilgi modelinde ise 2-5 terim sayısı ile en başarılı sonuçlara ulaşıldığını göstermektedir. TDT derlemi üzerinde yapılan test sonuçlarından farklı olarak, bu çalışmadan elde edilen sonuçlar, yöntem ne olursa olsun, uygun aralıktaki terim sayıları kullanılmadığı durumda (altında ya da üstünde terim sayısı) başarımın olumsuz etkilendiğini göstermektedir. Buna ek olarak, Türkçe bir derlem üzerinde konu izleme görevi ile ilgili olarak gerçekleştirilen başarım testlerinde, belge gösterimi için en uygun terim sayısı 60 olarak belirtilmesine rağmen, bunun altında ya da üstündeki değerlerin başarım üzerindeki etkileri konusunda bir yorum yapılmamıştır (Can ve diğerleri, 2010).

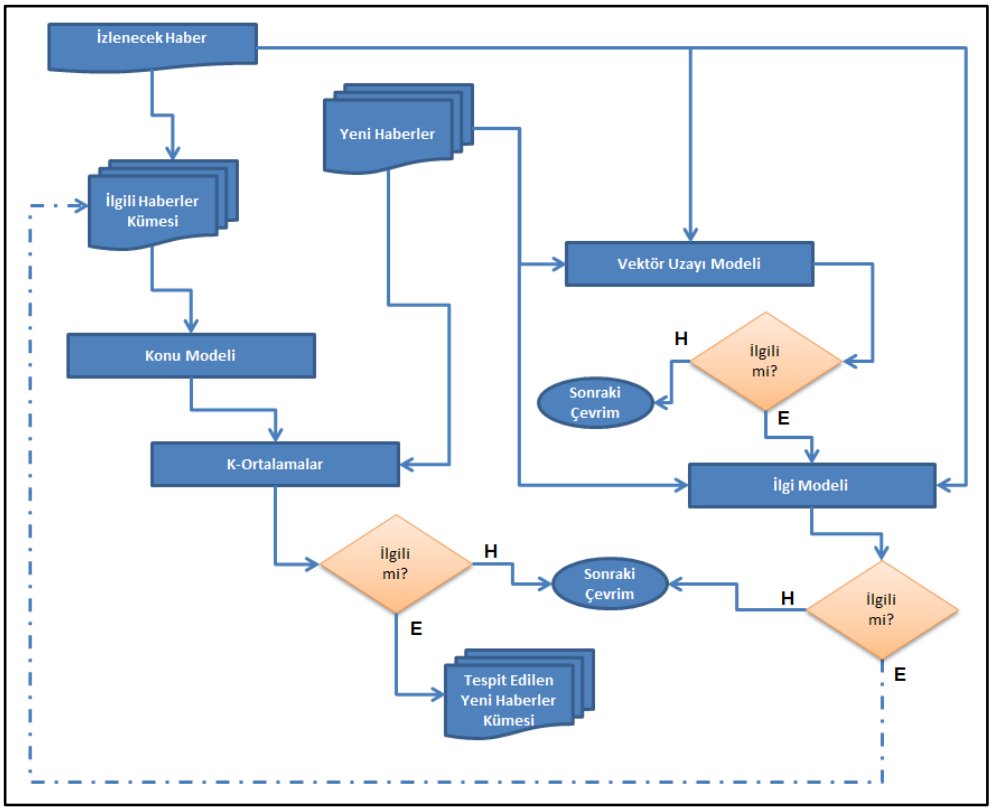
4.4. KONU İZLEME SİSTEMİ MİMARİ ÖNERİSİ

Bu çalışma kapsamında, gerek hikâye bağlantı algılama, gerekse konu izleme görevleri ile ilgili olarak gerçekleştirilen başarım testlerinin sonucunda, Türkçe için etkin bir konu izleme sistemi mimarisi oluşturabilmek için gerekli temel bileşenler belirlenmiştir. Etkin bir konu izleme sisteminde, izlemenin karşılaşılan ilk haberle birlikte başlayabilmesi, konu modelinin güçlü olması, konu modeli ile sisteme ulaşan yeni haberleri karşılaştıracak başarımı yüksek bir erişim fonksiyonu kullanılması ve konu modelini besleyecek duyarlığı yüksek bir geri besleme mekanizmasının tanımlanması, en kritik görevler olarak görülmektedir.

Akademik çalışmalarda, genellikle, izleme sürecine başlamadan önce konuyla ilgili birden fazla habere ihtiyaç duyulmakla birlikte, gerçek dünya uygulamalarında, bu süreç kullanıcının ilgisini çeken ilk haberle birlikte başlamaktadır. Bu nedenle, bir konu izleme sisteminin pratikte uygulanabilir olması için, izleme sürecini kullanıcının ilgilendiği ilk haberle birlikte başlatacak mimari yapı üzerine kurması gereklidir. Konu izleme sistemlerinde kritik işlemlerden bir tanesi de, güçlü konu modellerine sahip olma gereksinimidir. Konu izleme sistemleri, haberleri doğrudan karşılaştırmak yerine, sisteme ulaşan yeni haberleri, izlenen konuyu ifade ettiği varsayılan konu modelleri ile karşılaştırmayı tercih eder. Bu sayede izlenmek istenen konuların bağlamı çok daha iyi ifade edilebilmekte ve kelime kesişme olasılıkları artırılmaktadır. Konu izleme sistemlerinde, konu kümeleri genellikle, kümeleme yaklaşımı ile konu merkezi vektörleri oluşturularak gerçekleştirilmektedir. Bir konuyu ifade etmek için ilgili haberlerin seçilme yöntemleri son derece önemlidir. Bu noktada kullanıcılardan alınacak ilgililik geribildirimleri kullanılabileceği gibi, sistemi otomatik olarak beslemek için duyarlık başarımı yüksek yöntemlerin kullanılması da tercih edilebilir. Konu izleme sistemlerinde son aşamada ise, sisteme ulaşan yeni haberlerin izlenen konuyla ilgili olup olmadığına karar verecek olan erişim fonksiyonu yer alır. Bu tür bir sistemde, erişim fonksiyonu olarak kullanılacak yöntemin, pratikte kabul edilebilir ve hem anma hem de duyarlık başarımı olarak dengeli bir yapıya sahip olması gereklidir.

Bu çalışma kapsamında elde edilen sonuçlar, etkin bir konu izleme sistemi mimari önerisi açısından değerlendirildiğinde; hikâye bağlantı algılama tarafında duyarlığın yüksek olduğu vektör uzayı ve ilgi modellerinin AND birleşimleri, konu izleme tarafında ise f-ölçü başarımının yüksek olduğu k-ortalamalar yönteminin öne çıktığı görülmektedir. Tasarlanacak sistemde, izleme süreci karşılaşılan ilk haberle başlayacak olup, başlangıçta konu modeli sadece ilgili ilk haber kullanılarak oluşturulabilecektir. Bu durum, her ne kadar sistemin en zayıf tarafı gibi görünse de konu modelini besleyecek duyarlığı yüksek bir mekanizma sayesinde, sonraki aşamalarda konu modelinin güçlenmesi sağlanabilir. Diğer taraftan, konu modelini yeni gelen ilgili haberlerle besleyecek yapı ile yeni gelecek haberin izlenen konuya ait olup olmadığını belirlemeyi sağlayacak tespit yönteminin farklı seçilmesi, konu modelinin zaman içerisinde konu dışına çıkma olasılığını düşürmek açısından önemlidir. Bu sayede

sistemin verdiği ilgililik kararları ile dinamik olarak genişleyen bir konu modeli yerine, izlemenin başladığı ilk haberle ilgili olduğu güçlü kanıtlara dayandırılan haberlerle, konu modelinin kontrollü olarak genişlemesi sağlanabilecektir. Bu çalışmadan elde edilen sonuçlar ve etkili bir konu izleme sistemi için dikkat edilmesi gereken kıstaslar göz önüne alınarak tasarlanan izleme mimarisi, Şekil 8'de sunulmuştur. Şekilde, izleme süreci, kullanıcının ilgi alanında bulunan ilk haberin sisteme ulaşması ile başlamakta ve sonraki işleyişin hangi yönde ilerleyeceği oklarla ifade edilmektedir.



Şekil 8. Türkçe haberler için konu izleme sistemi mimarisi

Konu izleme sistemi işlevsel mimarisinde, kullanıcı ilgi alanındaki bir haber, zaman içerisinde izlenmek üzere belirlendikten sonraki ilk aşama, bu ilk haberin ilgili haberler kümesine aktarılmasıdır. Başlangıçta sistemde bu ilk haber dışında başka haber olmayacağı için, doğal olarak, ilk konu modelinin (ya da küme merkezi vektörünün) oluşturulması işlemi de bu ilk haber kullanılarak gerçekleştirilecektir. Aslında işlevsel mimaride izleme sürecinin başlatıldığı bu ilk çevrimde, konu modeli oluşturulması dışında yapılması gereken başka bir işlem yoktur.

İşlevsel mimaride sonraki çevrim, sisteme konuyla ilgili olup olmadığı bilinmeyen yeni haberlerin ulaşması ile başlar. Sisteme ulaşan bu yeni haber, bu çalışma kapsamında konu izleme görevinin gerçekleştirilmesinde en yüksek f-ölçü başarımına ulaşılan k-ortalamalar yöntemi kullanılarak, daha önceki aşamada oluşturulmuş olan konu modeli ile karşılaştırılır. Bu karşılaştırma sonucunda elde edilen benzerlik skoru, eğer daha önce belirlenmiş olan eşik değerinin üzerinde kalırsa, haber ilgili, altında kalırsa ilgisiz olarak işaretlenir. Haber ilgili bulunursa, tespit edilen yeni haberler kümesine aktarılır ve bir sonraki çevrime geçilir. İlgisiz bulunması durumunda ise, sistemde herhangi bir işlem yapılmaz ve bir sonraki çevrime geçilir.

Sisteme yeni bir haber ulaştığında, bu haberin ilgili haberler kümesinin bir elemanı olup olmayacağını belirlemek için, paralel olarak ikinci bir tespit işlemi başlatılır. Bu aşamada, sisteme ulaşan yeni haber, izleme işleminin başlatıldığı ilk haber kulanılarak, önce vektör uzayı modeli, sonra da ilgi modeli kullanılarak karşılaştırılır. Bu tespit gerçekleştirilirken, yeni haber için, eğer vektör uzayı modelinden ilgisiz kararı çıkarsa, bir sonraki tespit yöntemi olan ilgi modeli hiç çalıştırılmaz ve haber ilgisiz olarak işaretlenir. Sisteme yeni ulaşan haber için, hem vektör uzayı hem de ilgi modelinden konuyla ilgili kararı çıkması durumunda, bu yeni haber, ilgili haberler kümesine aktarılır ve konu modelinin yeniden oluşturularak güncellenmesi sağlanır.

Konu izleme sistemi ile ilgili olarak önerilen bu mimari yapı üzerinde bazı değişiklikler yapılarak, mimari yapının farklı ihtiyaçlar için özelleştirilmesi de mümkündür. Örneğin, güçlü konu modeli oluşturmak için ilgili haberler kümesine bir haber gönderilmeden önce, k-ortalamalar yöntemi de kullanılarak, ilgililik kanıtının çok daha kuvvetli olması sağlanabilir. Buna ek olarak, önerilen mimari tasarımın en büyük handikapı, üç farklı yöntemin gerçekleştirilmesine ihtiyaç duymasıdır. Her ne kadar sistem başarımını (anma ve duyarlık) en üst seviyede tutmak amacıyla yapılmış olsa da, gerçekleştirme maliyeti daha düşük ve daha hızlı bir sisteme ihtiyaç duyulması durumunda, sadece k-ortalamalar yöntemi kullanılarak da bir izleme sistemi gerçekleştirilmesi mümkündür. Bunu gerçekleştirmek için, tasarımdan vektör uzayı ve ilgi modelleri modüllerini çıkararak, ilgili haberler kümesini beslemek için sadece k-ortalamalar yönteminin kararına güvenmek yeterli olacaktır.

Bunlara ek olarak, yöntemlerin eşik değerlerinin ve belge gösteriminde uygun terim sayılarının dinamik olarak belirlenebilmesi için, son kullanıcı geribildirimlerinden yararlanacak bir mekanizma tanımlanması mümkündür. Son kullanıcı kararlarının tasarım içerisinde yer alması ile ilgililik değerlendirmesi için en güçlü kanıt elde edilmiş olur ve gerek konu modelinin güncellenmesi, gerekse yöntemlerin eşik değerlerinin saptanmasında, bu kararlar sistem parametrelerinin dinamik olarak belirlenmesinde kullanılabilir.

Önerilen mimari yapı, bu çalışma kapsamında elde edilen somut başarımlar göz önünde bulundurularak oluşturulmuştur. Mimarinin bileşenleri pratikte uygulanabilir yöntemler olarak belirlenmiş olup, yapı üzerinde küçük değişikliklerle, farklı bilgi erişim ihtiyaçlarına rahatlıkla cevap verebilecek esnekliktedir.

5. BÖLÜM

SONUÇ

5.1. SONUÇLAR

TDT programı içerisinde tanımlanmış olan hikâye bağlantı algılama ve konu izleme görevleri için erişim fonksiyonu ve belge gösterimi tarafında farklı yöntemler kullanarak, Türkçe bir derlem için erişim başarımının artırılmasını ve etkin bir konu izleme mimarisi ortaya konulmasını amaçlayan bu çalışmada, belirlenen amacı gerçekleştirebilmek için BilCol-2005 derlemi üzerinde bir dizi test uygulanmıştır. Bu bölümde, gerçekleştirilen başarım testlerinden elde edilen sonuçlar yorumlanarak, bu sonuçların Türkçe bir bilgi erişim sistemine ne tür katkılar sağlayacağı sunulmakta ve araştırma için belirlenen hipotezlerin doğrulanıp doğrulanmadığı ortaya konulmaktadır.

Uygulanan yöntemlerin başarım değerlendirmeleri f-ölçü, anma ve duyarlık değerlerine göre yorumlanmıştır. Anma ve duyarlık değerlerinin harmonik ortalamalarını veren f-ölçü değeri, yöntemlerin genel başarımlarının değerlendirilmesi ve sonuçların karşılaştırılmasında son derece kullanışlı bir metriktir. Diğer taraftan, bilgi erişim sistemleri tasarlanırken, sistemlerin başarımı, gerçekleştirilecek sistemden ne beklendiği ile de doğrudan ilgilidir. Bu kapsamda tasarlanan sistem, ilgili belgelerin büyük bir çoğunluğuna erişmeyi hedefliyorsa, başarım metriğinin anma, erişilen belgelerin büyük çoğunluğunun ilgili olması isteniyorsa, başarım metriğinin duyarlık olarak belirlenmesi daha anlamlıdır. Bu bağlamda, başarım testleri her üç kıstas da göz önünde bulundurularak değerlendirilmektedir.

Hikâye bağlantı algılama görevi ile ilgili olarak gerçekleştirilen testlerde, Türkçe bir derlem üzerinde, bu görevin gerçekleştirilmesinde, erişim fonksiyonu olarak vektör uzayı ve ilgi modellerinin kullanılmasının, erişim başarımı üzerindeki etkisi ortaya konulmuştur. Gerçekleştirilen başarım testlerinden elde edilen sonuçlar, hikâye bağlantı algılama görevinde erişim fonksiyonu olarak, vektör uzayı modelinin kullanılmasının, ilgi modeli kullanımına göre çok daha yüksek f-ölçü, anma ve duyarlık başarımı elde edilmesini sağladığını göstermektedir. Ulaşılan bu sonuç, "*hikâye bağlantı algılama*

görevinde erişim fonksiyonu olarak ilgi modelinin kullanılması, erişim fonksiyonu olarak vektör uzayı modeli kullanılmasına göre daha yüksek f-ölçü değeri sağlar" hipotezinin doğrulanamaması anlamına gelmektedir.

İlgi modelinden elde edilen başarım değerlerinin, vektör uzayı modeline göre daha düşük olması, üzerinde tartışılması gereken bir konudur. Literatürde, özellikle hikâye bağlantı algılama görevinin gerçekleştirilmesinde uygulanan ilgi modelinin başarımı, genellikle vektör uzayı modelinden daha yüksek olarak gösterilmektedir (Connell ve diğerleri, 2004; Lavrenko ve diğerleri, 2002). Bu çalışmada, ilgi modeli için elde edilen başarım değerlerinin düşük kalmasının en önemli nedeninin, uygulanan yöntemdeki ön işlemlerden kaynaklandığı düşünülmektedir. Bilindiği gibi, ilgi modelinde her bir belge için konu modelleri yaratılmakta ve belge benzerlikleri hesaplanırken oluşturulan bu konu modelleri doğrudan karşılaştırılmaktadır. Doğal olarak belgeler için oluşturulan konu modelleri ne kadar güçlü olursa, belgeleri birbiri ile eşleştirmek de o kadar kolay olacaktır. Bu kapsamda, literatürdeki çalışmalarda, konu modelleri oluşturulurken, başlangıçta her bir belgedeki terimler sorgu olarak kabul edilmekte, bu sorgu eğitim kümesine yollanarak en ilgili bulunan ilk *N* adet belge alınmakta ve konu modeli sadece belgeye göre değil, elde edilen bu belge listesine göre oluşturulmaktadır. Bu yaklaşım, doğal olarak belgeler için oluşturulan konu modellerindeki terimlerin daha sağlam kanıtlarla seçilmesine olanak tanımakta ve oluşturulan konu modelleri daha güçlü olmaktadır. Diğer taraftan, bu yaklaşım, elimizde iyi oluşturulmuş bir eğitim belge kümesi olmasını gerektirir ki, bu ihtiyaç bu yaklaşımın en büyük dezavantajını oluşturur. Gerçek zamanlı çalışacak bir bilgi erişim sisteminde, böyle bir kümeye erişimin olamayacağı ya da çok zor olacağı açıktır. Bu nedenle, bu çalışmadan elde edilecek sonuçların pratikte uygulanabilir sistemlere temel olması amaçlandığından, literatürde uygulanan bu yöntem benimsenmemiştir. Bunun yerine her bir belge için konu modelleri, belgede geçen terimlerin belgeyi temsil etme olasılıklarına göre hesaplanmış ve belgeler, bu yaklaşımla oluşturulan konu modelleri temel alınarak karşılaştırılmıştır. Bu kapsamda, ilgi modeli için elde edilen başarım değerlerinin literatürdeki çalışmalardan düşük çıkması çok şaşırtıcı değildir.

Hikâye bağlantı algılama görevinde farklı erişim fonksiyonları kullanılırken, belgeleri göstermek için kullanılan terim sayılarının, erişim başarımı üzerinde önemli etkileri

olduğu tespit edilmiştir. F-ölçü başarım değerlerine göre, vektör uzayı modelinde belgeleri göstermek için kullanılması gereken en uygun terim sayısı *tf.idf* değeri en yüksek ilk 30 terimken, ilgi modelinde bu değer 4 olarak tespit edilmiştir. Erişim fonksiyonunda farklı yöntemler kullanılırken, belge gösterimi için uygun değerlerin seçilmemesi halinde, vektör uzayı modelinde %15'e, ilgi modelinde ise %10'a kadar f-ölçü başarım değişimleri gözlenmiştir.

Buna ek olarak, vektör uzayı modelinde, belgeleri göstermek için 10 ile 40 arasında seçilecek terim sayıları, f-ölçü başarımı üzerinde önemli bir değişim yaratmamaktadır. Başarım, 30 terimde en yüksek seviyesine ulaşmakta, sonrasında ise terim sayısının artmasına paralel olarak küçük düşüşler yaşanmaktadır. İlgi modelinde ise, belgeleri göstermek için seçilen terim sayısı, 2-5 aralığında en yüksek f-ölçü başarımına ulaşmıştır. Diğer taraftan, ilgi modelinde seçilen terim sayısı 5'in üzerine çıktıktan sonra, terim sayısı arttıkça f-ölçü başarımının düştüğü belirlenmiştir. Belge gösterimi ile ilgili olarak elde edilen bu bulgular "*hikâye bağlantı algılama görevinde belgeleri ifade etmek için kullanılan terim sayısı arttıkça f-ölçü başarımı da artar*" hipotezinin doğrulanamadığı anlamına gelmektedir.

Belge gösterimi için elde edilen bulgular, özellikle vektör uzayı yöntemi göz önüne alındığında, Can ve diğerleri (2010) tarafından gerçekleştirilen çalışmada en yüksek başarımın elde edildiği 60 terim sayısından daha az terim sayısını işaret etmektedir. Diğer taraftan tamamen İngilizce belgelerle oluşturulmuş olan TDT derlemi üzerinde yapılan çalışma (Lavrenko ve diğerleri, 2002), 30 terim veya üzerinde terim sayısının en yüksek başarım için yeterli olduğunu göstermektedir. Bu çalışmadan elde edilen bulgular ışığında, Türkçe derlemler üzerinde erişim fonksiyonu olarak kullanılan yönteme göre, belgeleri göstermek için en yüksek f-ölçü başarımına ulaşılmasını sağlayan terim sayısı ya da aralığının belirlenmesinin son derece önemli olduğu söylenebilir. Bu aralığın altında ya da üstünde terim sayılarının seçilmesi, başarımı olumsuz olarak etkilemektedir.

Hikâye bağlantı algılama görevi için erişim fonksiyonu olarak kullanılan vektör uzayı ve ilgi modellerinin, AND mantıksal birleşimi ile elde edilen sonuçlara bakıldığında ise, beklendiği gibi anma değerlerinin düştüğü, duyarlık değerlerinin ise yükseldiği

görülmektedir. Buna göre yöntemlerin AND birleşimlerindeki duyarlık değerlerinde, aynı terim sayıları için, vektör uzayı modelinde %50'lerin, ilgi modelinde ise %60'ların üzerinde artış tespit edilmiştir. Bu bulgular ışığında *"hikâye bağlantı algılama görevinde erişim fonksiyonu olarak vektör uzayı modeli ve ilgi modelinin AND birleşimlerinin kullanılması, modellerin tek başlarına kullanıldığı yaklaşıma göre daha yüksek duyarlık değeri sağlar"* hipotezi doğrulanmıştır.

Bilgi erişim sistemlerinde, farklı erişim fonksiyonları için, AND birleşimleri genellikle daha seçici sistemler tasarlamak için kullanılmaktadır. Bu tür sistemlerin eriştikleri belgelerin büyük bir bölümü ilgili olmakla birlikte, toplam erişilen ilgili belge sayısı, yöntemlerin bağımsız olarak uygulandığı duruma göre daha düşük kalmaktadır. Bu bağlamda, hikâye bağlantı algılama görevi için erişim fonksiyonu olarak kullanılan vektör uzayı ve ilgi modellerinin AND birleşimlerinden elde edilen sonuçlar, duyarlık başarımı açısından yöntemlerin bağımsız olarak uygulandığı duruma göre önemli oranda artış sağlamıştır. Buna ek olarak, yöntemlerin AND birleşimlerinde, belgeleri göstermek için kullanılan terim sayısı arttıkça duyarlık başarımlarının da paralel olarak artış gösterdiği gözlenmiştir.

Hikâye bağlantı algılama görevi için erişim fonksiyonu olarak kullanılan vektör uzayı ve ilgi modellerinin, OR mantıksal birleşimi ile elde edilen sonuçlarına bakıldığında ise, beklendiği gibi, genel olarak duyarlık değerlerinin düştüğü, anma değerlerinin ise yükseldiği görülmektedir. Buna göre, yöntemlerin OR birleşimlerindeki anma değerlerinde, aynı terim sayıları için, vektör uzayı modelinde %5'lerin, ilgi modelinde ise %20'lerin üzerinde artış tespit edilmiştir. Bu bulgular ışığında, *"hikâye bağlantı algılama görevinde erişim fonksiyonu olarak vektör uzayı modeli ve ilgi modelinin OR birleşimlerinin kullanılması, modellerin tek başlarına kullanıldığı yaklaşıma göre daha yüksek anma değeri sağlar"* hipotezi doğrulanmıştır.

Bilgi erişim sistemlerinde farklı erişim fonksiyonları için OR birleşimleri genellikle daha çok ilgili belgeye erişecek ya da ilgili belgeleri kaçırmayacak sistemler tasarlamak için kullanılmaktadır. Bu tür sistemler, ilgili belgelerin büyük bir bölümüne erişirken, erişim çıktısında yer alan belgelerin ilgisiz olma olasılığını da artırır. Bu bağlamda, hikâye bağlantı algılama görevi için erişim fonksiyonu olarak kullanılan vektör uzayı ve

ilgi modellerinin, OR birleşimlerinden elde edilen sonuçlar anma başarımı açısından yöntemlerin bağımsız olarak uygulandığı duruma göre, sınırlı bir oranda artış sağlamıştır. Diğer taraftan, f-ölçü başarımı olarak değerlendirildiğinde, OR birleşiminden elde edilen sonuçlar vektör uzayı modelinden elde edilen başarımın altında kalmaktadır. Buna ek olarak, yöntemlerin OR birleşimlerinde belgeleri göstermek için kullanılan terim sayısı arttıkça, ilgi modelinde genellikle anma başarımı da artmış, ancak benzer bir ilişki, vektör uzayı modeli için tespit edilememiştir.

Hikâye bağlantı algılama görevi ile ilgili olarak elde edilen diğer bir ilginç sonuç ise, eşik değeri ile ilgilidir. Buna göre, vektör uzayı modelinde 50 terime kadar, belgeleri göstermek için kullanılan terim sayılarındaki değişiklilikler eşik değerlerinde önemli oranda farklılıklar göstermişken, 50 terimden sonra eşik değerinin çok daha küçük oranlarda değiştiği ya da aynı kaldığı gözlenmiştir. Bu kapsamda, belgeleri göstermek için seçilecek terim sayısının 50 ya da altında olduğu durumlarda, eşik değerlerindeki küçük oynamaların başarım üzerinde ciddi farklar yaratacağı söylenebilir. Bu nedenle, vektör uzayı modeli tabanlı bir bilgi erişim sisteminde, belgeleri göstermek için seçilecek terim sayısı belirlendikten sonra, eşik değerinin en uygun değerinin belirlenmesi, en yüksek başarımın elde edilebilmesi açısından son derece önemlidir. Buna karşılık, vektör uzayı modelinden farklı olarak, ilgi modelinde, belgeleri göstermek için kullanılan terim sayıları ile eşik değerler arasında doğrusal bir ilişki tespit edilememiştir.

Bu çalışma kapsamında, konu izleme görevi ile ilgili olarak k-ortalamalar, vektör uzayı ve ilgi modeli yöntemleri kullanılarak başarım testleri gerçekleştirilmiştir. K-ortalamalar yöntemi, kümeleme tabanlı bir yöntemdir ve bu yöntem uygulanmadan önce uygun eşik değeri belirleme yöntemini belirleyebilmek için bazı testler uygulanmıştır. Uygulanan yöntemler arasındaki başarım farkları, eşik değeri belirleme yöntemlerinin sistem başarımları üzerinde ciddi etkisi olduğunu göstermektedir. Elde edilen sonuçlar, hikâye bağlantı algılama görevinde de kullanılan, eğitim kümesinde anma-duyarlık değerinin birlikte en yüksek olduğu noktanın seçilmesi yöntemini öne çıkarmıştır. Bu yöntemle, başarım olarak kendisine en yakın olan küme konularına göre, eşik değerinin dinamik olarak belirlendiği yönteme göre %36,71 oranında daha başarılı sonuçlar elde edilmiştir. Elde edilen bu bulgular, *"konu izleme görevinde kümeleme için eşik değeri*

olarak 'anma ve duyarlığın en yüksek olduğu değerin seçildiği yöntemin' kullanılması 'küme merkezi vektörüne eğitim belgelerinin uzaklığını temel alan yöntemlere' göre daha yüksek f-ölçü başarımı elde edilmesini sağlar" hipotezini doğrulamıştır.

Bununla birlikte anma ve duyarlığın birlikte en yüksek olduğu noktaya göre eşik değeri seçme işlemi, gerçek sistemler üzerinde pratikte uygulanması zor olan bir yöntemdir. Bu yöntemle uygun eşik değerinin belirlenebilmesi için, ilgililik değerlendirmesi yapılmış ve kabul edilebilir sayıda eğitim belgesi gerekmektedir. Konu izleme sistemleri için, başlangıçta, eşik değerini belirleyecek sayıda eğitim belgesi bulmak kolay olmadığı için, yöntemin pratikte uygulanması problemlidir. Bu nedenle, her ne kadar başarımı daha düşük olsa da, gerçek zamanlı sistemlerde, eşik değerlerinin her bir konu için ayrı ayrı dinamik olarak belirlendiği yöntem eşik belirleme yöntemi olarak kullanılabilir.

Konu izleme ile ilgili olarak uygulanan yöntemlerin başarımları f-ölçü değerlerine göre karşılaştırıldığında ise, k-ortalamalar yöntemi öne çıkmaktadır. K-ortalamalar yönteminden elde edilen f-ölçü başarım değeri, hem vektör uzayı modelinden hem de ilgi modelinden daha yüksektir. Bu bağlamda, *"konu izleme görevinde erişim fonksiyonu olarak kümeleme tabanlı bir yöntemin kullanılması vektör uzayı ya da ilgi modelinin kullanıldığı yönteme göre daha yüksek f-ölçü başarımı elde edilmesini sağlar"* hipotezi doğrulanmıştır.

Genel olarak, yöntemlerin konu izleme görevi için peformanslarına bakıldığında, k-ortalamalar yönteminin öne çıktığı görülmekle birlikte, k-ortalamalar uygulamada zaman maliyeti daha yüksek bir yöntemdir. Bir bilgi erişim sistemi tasarlanırken, birden fazla erişim yönteminin kullanılması, hem zaman maliyetini hem de ihtiyaç duyulacak donanım kaynaklarının sayısını artıracaktır. Bu nedenle, böyle bir sistem tasarlanırken performans–maliyet dengesinin iyi ayarlanması gereklidir. Bu kapsamda başarımın önemli olduğu sistemlerde, k-ortalamalar, maliyetin önemli olduğu sistemlerde de vektör uzayı yönteminin kullanılmasının daha uygun olacağı düşünülmektedir.

Tüm bu sonuçlara ek olarak, bu çalışma kapsamında uygulanan yöntemler ve elde edilen başarım ölçümleri değerlendirildiğinde, Türkçe haberler için konu izleme ile

ilgili olarak, pratikte uygulanabilecek bir bilgi erişim sisteminin de temel bileşenleri belirlenmiştir. Bu kapsamda, güçlü konu modelleri oluşturabilmek için en uygun yöntemin vektör uzayı modeli ile ilgi modelinin AND birleşimi olduğu, güçlü konu modelleri oluşturulduktan sonra da, k-ortalamalar algoritması kullanılarak yeni gelen belgelerin takip edilen konuyla ilgili olup olmadıklarının belirlenebileceği düşünülmektedir. Böylece konu modellerinin oluşturulmasından başlayarak, konuyla ilgili yeni gelen belgelerin tespit edilmesine kadar başarımı yüksek yöntemlerin kullanılması sağlanmış olacaktır.

Bu çalışma kapsamında, TDT programında tanımlanmış olan hikâye bağlantı algılama ve konu izleme alt görevlerinin, Türkçe bir derlem üzerinde gerçekleştirilmesi için belge gösterimi ve erişim fonksiyonu bacaklarında çeşitli testler uygulanmış ve sonuçlar ortaya konulmuştur. Sonraki aşamada, başarım testlerinden elde edilen somut sonuçlar değerlendirilerek, Türkçe için etkin ve esnek bir konu izleme sistemi mimarisi önerilmiştir.

5.2. GELECEK ÇALIŞMALAR

Bu çalışmanın devamında şu konularda ek çalışmalar yapılmasının anlamlı olacağı düşünülmektedir:

- Hem hikâye bağlantı algılama hem de konu izleme görevleri için eşik değeri belirlemek son derece kritiktir. Bu kapsamda, her iki görev için de bu çalışmada gerçekleştirilen testlere benzer, ancak çok daha kapsamlı çalışmaların yapılması gerekmektedir.

- Belge gösterimi tarafında farklı yaklaşımlar test edilerek sonuçların irdelenmesi gerekmektedir. Özellikle varlık isimlerinin kullanılması ve olay modeli (event model) tabanlı yaklaşımların Türkçe derlemler üzerinde denenerek sonuçlarının tartışılması gereklidir.

- Bu çalışma, hikâye bağlantı algılama ve konu izleme görevlerinde ağırlıklı olarak erişim başarımı üzerine yoğunlaşmıştır. Oysa pratikte bu tür sistemleri

gerçekleştirirken kullanılacak yöntemlerin, zaman ve gerçekleştirme maliyetleri de önemli bir bileşen olmaktadır. Bu kapsamda, erişim başarımı yüksek yöntemlerin zaman maliyeti değerlendirmeleri yapılarak, maliyeti düşük ve başarımı yüksek yöntemlerin ortaya konulması gereklidir,

- TDT konusunda Türkçe derlemler üzerindeki çalışmalar son derece sınırlı olup, bu program içerisinde tanımlı farklı görevler için Türkçe derlemler üzerindeki çalışmaların sayısının artırılması gerekir.

Gelecekte yapılması önerilen tüm bu çalışmaların yanında, bu çalışma kapsamında önerilen konu izleme işlevsel mimari yapısının, gerek başarım gerekse zaman maliyeti açısından, gerçek sistemler üzerinde denenmesi ve sonuçların görülmesi gerekmektedir.

KAYNAKÇA

Acun, B., Başpınar, A., Oğuz, E., Saraç, M. İ. ve Can, F. (2013). Topic tracking using chronological term ranking. Erol Gelenbe ve Ricardo Lent (Ed.), *Computer and Information Sciences III* içinde (s. 353-361). London: Springer.

Aksoy, C., Can, F. ve Koçberber, S. (2012). Novelty detection for topic tracking. *Journal of the American Society for Information Science and Technology*, *63*(4): 777-795.

Allan, J. (2002). Introduction to topic detection and tracking. James Allan (Ed.), *Topic Detection and Tracking* içinde (s. 1-16). USA: Springer.

Allan, J., Lavrenko, V. ve Jin, H. (2000). First story detection in TDT is hard. *Proceedings of the 2000 ACM CIKM International Conference on Information and Knowledge Management, McLean, VA, USA, November 6-11, 2000* içinde (s. 374-381). ACM 2000.

Allan, J., Lavrenko, V. ve Swan, R. (2002). Explorations within topic detection and tracking. James Allan (Ed.), *Topic Detection and Tracking* içinde (s.197-224). USA: Springer.

Allan, J., Carbonell, J., Doddington, G., Yamron, J. ve Yang, Y. (1998). Topic detection and tracking pilot study: final report. *Proceedings of the Broadcast News Transcription and Understanding Workshop, February 8-11, 1998, Lansdowne Conference Resort, Lansdowne, Virginia* içinde (s. 194-218).

Allan, J., Lavrenko, V., Frey, D. ve Khandelwal, V. (2000). UMass at TDT 2000. *Proceedings of Topic Detection and Tracking Workshop* içinde (s. 109-115). USA: National Institute of Standard and Technology.

Bağlıoğlu, Ö. (2009). *New event detection using chronological term ranking.* Yayımlanmamış Yüksek Lisans Tezi, Bilkent Üniversitesi Fen Bilimleri Enstitüsü. 10 Mayıs 2014 tarihinde http://www.thesis.bilkent.edu.tr/0006330.pdf adresinden erişildi.

Balabantaray, R. C., Swain, M. ve Sahoo, B. (2013). Evaluation of web search engines based on ranking of results and features. *International Journal of Human Computer Interaction* (IJHCI), *4*(3), 117-127.

Belkin, N.J., Kantor, P., Fox, E.A. ve Shaw, J.A. (1995). Combining the evidence of multiple query representations for information retrieval. *Information Processing & Management, 31*, 431-448.

Berger, A. ve Lafferty, J. (1999). Information retrieval as statistical translation. *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 15-19, 1999, Berkeley, CA, USA* içinde (s. 222-229). New York, NY: ACM.

Berry, M. W. ve Castellanos, M. (Eds.). (2004). *Survey of text mining*. New York: Springer.

Bhogal, J., Macfarlane, A. ve Smith, P. (2007). A review of ontology based query expansion. *Information Processing & Management, 43*(4), 866-886.

Bikel, D. M., Schwartz, R. ve Weischedel, R. M. (1999). An algorithm that learns what's in a name. *Machine Learning, 34*(1-3), 211-231.

Braschler, M. ve Ripplinger, B. (2004). How effective is stemming and decompounding for German text retrieval?. *Information Retrieval, 7*(3-4), 291-316.

Bun, K. K. ve Ishizuka, M. (2001). Emerging topic tracking system. Ning Zhong, Yi Yu Yao, Jiming Liu ve Setsuo Ohsuga (Ed.), *Web Intelligence: Research and Development, First Asia-Pacific Conference, WI 2001, Maebashi City, Japan, October 23-26, 2001, Proceedings* içinde. (s. 125-130). Springer-Verlag, London, UK.

Bun, K. K. ve Ishizuka, M. (2006). Emerging topic tracking system in WWW. *Knowledge-Based Systems*, *19*(3), 164-171.

Can, F., Altingövde, I. S. ve Demir, E. (2004). Efficiency and effectiveness of query processing in cluster-based retrieval. *Information Systems*, *29*(8), 697-717.

Can, F., Kocberber, S., Baglioglu, O., Kardas, S., Ocalan, H. C. ve Uyar, E. (2010). New event detection and topic tracking in Turkish. *Journal of the American Society for Information Science and Technology*, *61*(4), 802-819.

Can, F., Kocberber, S., Baglioglu, O., Kardas, S., Ocalan, H. C. ve Uyar, E. (2008). Bilkent News Portal: A personalizable system with new event detection and tracking capabilities. S.-H. Myaeng, D. W. Oard, F. Sebastiani, T.-S. Chua ve M.-K. Leong (Ed.), *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, July 20-24, 2008, Singapore* içinde (s. 885-885). Singapore: ACM.

Can, F., Koçberber S., Bağlıoğlu, O., Kardaş, S., Öcalan, H. C. ve Uyar, E. (2007). Türkçe haberlerde yeni olay bulma ve izleme: Bir deney derleminin oluşturulması. Serap Kurbanoğlu, Umut Al, Phyllis Lepon Erdoğan, Yaşar Tonta ve Nazan Özenç Uçak (Ed.), *3rd International Symposium on Information Management in a Changing World, September 19-21, 2012, Ankara, Turkey* içinde (s.50-59). Ankara: Hacettepe Üniversitesi Bilgi ve Belge Yönetimi.

Carthy, J. ve Sherwood-Smith, M. (2002). Lexical chains for topic tracking. Abdelkader El Kamel, Khaled Mellouli ve Pierre Borne (Ed.), *2002 IEEE International Conference on Systems, Man and Cybernetics, October 6-9,2002, Yasmine Hammamet - Tunisia* içinde (s.370-374). IEEE.

Carthy, J. ve Smeaton, A. (2000). The design of a topic tracking system. *Proceedings of the BCS-IRSG 22nd Annual Colloquium onInformation Retrieval Research: 5th-7th April 2000, SidneySussex College, Cambridge, England*. 19 Mayıs 2014 tarihinde http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.21.1493&rep=rep1&type=pdf adresinden erişildi.

Castells, P., Fernandez, M. ve Vallet, D. (2007). An adaptation of the vector-space model for ontology-based information retrieval. *Knowledge and Data Engineering, IEEE Transactions*, *19*(2), 261-272.

Chen, F., Farahat, A. ve Brants, T. (2004). Multiple similarity measures and source-pair information in story link detection. Julia Hirschberg (Ed.), *Human Language Technology Conference, North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT/NAACL 2004), May 2-7 2004, Boston; MA; USA* içinde (s. 313-320). East Stroudsburg, PA: ACL.

Chen, H., Yim, T., Fye, D. ve Schatz, B. (1995). Automatic thesaurus generation for an electronic community system. *Journal of the American Society for Information Science*, *46*(3), 175–193.

Chen, Y. M., Wang, X. L. ve Liu, B. Q. (2005). Multi-document summarization based on lexical chains. *Proceedings of 2005 International Conference on Machine Learning and Cybernetics: August 18-21, 2005, Guangzhou, China* içinde (s. 1937-1942). Piscataway, NJ: IEEE.

Connell, M., Feng, A., Kumaran, G., Raghavan, H., Shah, C. ve Allan, J. (2004). UMass at TDT 2004. *In Topic Detection and Tracking Workshop Report*. 19 Mayıs 2014 tarihinde http://comminfo.rutgers.edu/~chirags/papers/umass_tdt04.pdf adresinden erişildi.

Dai, X. Y., Chen, Q. C., Wang, X. L. ve Xu, J. (2010). Online topic detection and tracking of financial news based on hierarchical clustering. *International Conference on Machine Learning and Cybernetics, ICMLC 2010, July 11-14, 2010, Qingdao, China* içinde, (s. 3341-3346). Piscataway, N.J. : IEEE.

Diao, H., Bai, Z. ve Yu, X. (2010). The application of improved K-Nearest Neighbor classification in topic tracking. *International Conference on Educational and Information Technology (ICEIT), 17-19 Sept. 2010, Chongqing* içinde, (s. 64-68). IEEE.

Dinçer, B.T. (2004). *Türkçe için istatistiksel bir bilgi geri-getirim sistemi.* Yayımlanmamış doktora tezi. Ege Üniversitesi Fen Bilimleri Enstitüsü, 11 Mayıs 2014 tarihinde http://yunus.hacettepe.edu.tr/~tonta/courses/spring2011/bby704/B-TanerDincer%20-%20Doktora%20Tezi.pdf adresinden erişildi.

Ekmekçioglu, F. C. ve Willett, P. (2000). Effectiveness of stemming for Turkish text retrieval. *Program, 34*(2), 195-200.

Figuerola, C.G., Gomez, R., Rodriguez, A.F.Z. ve Berrocal, J.L.A. (2002). Stemming in Spanish: a first approach to its impact on information retrieval. Carol Peters (Ed.), *Working Notes for the CLEF 2001 Workshop, 3 September 2001, Darmstadt, Germany* içinde (s. 197-202). Springer.

Fiscus, J., Doddington, G., Garofolo, J. ve Martin, A. (1999). NIST's 1998 Topic Detection and Tracking evaluation (TDT2). Lynette Hirschman (Ed.), *Proceedings of the 1999 DARPA Broadcast News Workshop, February 28-March 3, 1999, Herndon, Virginia* içinde (s. 19-24). San Francisco, CA: Morgan Kaufmann Publishers.

Frakes, W. ve Baeza Yates, R. (1992). *Information retrieval: Data structure and algorithm, clustering algorithms*, Englewood Cliffs, NJ: Prentice-Hall.

Franz, M., McCarley, J. S., Ward, T. ve Zhu, W.-J. (2001). Unsupervised and supervised clustering for topic tracking. Donald H. Kraft, W. Bruce Croft, David J. Harper ve Justin Zobel (Ed.), *SIGIR '01 24th ACM/SIGIR International Conference on Research and Development in Information Retrieval, September 09 - 12, 2001, New Orleans, LA, USA* içinde (s. 310-317). New York: Association for Computing Machinery.

Gaines, B.R., Chen, L.L. ve Shaw, M.L.G. (1997). Modeling the human factors of scholarly communities supported through the Internet and World Wide Web. *Journal of the American Society for Information Science, 48*(11), 987–1003.

Gordon, M. ve Pathak, P. (1999). Finding information on the World Wide Web: The retrieval effectiveness of search engines. *Information Processing & Management, 35,* 141–180.

Gupta, V. ve Lehal, G.S. (2009). A survey of text mining techniques and applications. *Journal of Emerging Technologies in Web Intelligence*, 1(1), 60-76.

Han, H., Jeong, W. ve Wolfram, D. (2014). Log analysis of academic digital library: user query patterns. Kindling, M. ve Greifeneder, E. (Ed.), *iConference 2014 Proceedings, March 4-7, 2014, Berlin, Germany* içinde (s. 1002 - 1008). Illinois: iSchools.

Han, J. (1996). Data mining techniques. *ACM SIGMOD Record*, 25(2): 545.

Han, J., Kamber, M. ve Pei, J. (2006). *Data mining: concepts and techniques*. San Francisco, Calif. : Morgan Kaufmann.

Harman, D. (1991). How effective is suffixing? *Journal of the American Society for Information Science*, 42(1): 7-15.

Hatch, P., Stokes, N. ve Carthy, J. (2000). Topic detection, a new application for lexical chaining. *Proceedings of the BCS-IRSG 22nd Annual Colloquium onInformation Retrieval Research, 5th-7th April 2000, Cambridge, England.* içinde (s. 94-103). Swindon, Wiltshire: British Computer Society.

Hatzivassiloglou, V., Gravano, L. ve Maganti, A. (2000). An investigation of linguistic features and clustering algorithms for topical document clustering. Nicholas J. Belkin, Peter Ingwersen ve Mun-Kew Leong (Ed.), *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval: ACM SIGIR, July 24-28, 2000, Athens, Greece* içinde (s. 224-231). New York, N.Y. : Association for Computing Machinery.

Hearst, M. A. (2011). Natural search user interfaces. *Communications of the ACM*, 54(11), 60-67.

Hull, D. (1996). Stemming algorithms: a case study for detailed evaluation. *Journal of the American Society for Information Science*, 47(1), 70-84.

İlhan, U. (2001). *Application of k-nn and fptc based text categorization algorithms to Turkish news reports.* Yayımlanmamış Yüksek Lisans Tezi. Bilkent Üniversitesi Fen Bilimleri Enstitüsü. 10 Mayıs 2014 tarihinde http://www.cs.bilkent.edu.tr/tech-reports/2001/BU-CE-0104.pdf adresinden erişildi.

Jansen, B. J., Spink, A. ve Saracevic, T. (2000). Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing & Management*, *36*(2), 207-227.

Jin, Y., Myaeng, S.H., Lee, M., Oh, H. ve Jang, M. (2005). Effective Use of Place Information for Event Tracking. *Lecture Notes in Computer Science, 3689*, 410-422.

Kardaş, S. (2009). *New event detection and tracking in Turkish*. Yayımlanmamış Yüksek Lisans Tezi. Bilkent Üniversitesi Fen Bilimleri Enstitüsü. 10 Mayıs 2014 tarihinde http://www.thesis.bilkent.edu.tr/0003828.pdf adresinden erişildi.

Kaur, K. ve Gupta, V. (2012). A survey of topic tracking techniques. *International Journal of Advanced Research in Computer Science and Software Engineering*, 2(5), 384-393.

Khachumov, M. V. (2012). Distances, metrics and cluster analysis. *Scientific and Technical Information Processing*, 39(6), 310-316.

Kim, P. ve Myaeng, S.H. (2004). Usefulness of temporal information automatically extracted from news articles for topic tracking. *ACM Transactions on Asian Language Information Processing (TALIP),* 3(4), 227-242.

Köksal, A. (1981). Tümüyle özdevimli deneysel bir belge dizinleme ve erişim dizgesi: TÜRDER. *3. Ulusal Bilişim Kurultayı, 1981, Ankara* içinde (s. 37-44). Ankara: TBD.

Köse, G., Tonta, Y., Ahmadlouei, H. ve Polatkan, A. C. (2013). Story link detection in Turkish corpus. *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences, 17-20 Nov. 2013 içinde* (s. 154-158). Atlanta:IEEE.

Köse, G. ve Ahmadlouei, H. (2013). Supervised news classification based on a large-scale news corpus. *4th International Symposium on Information Management in a Changing World, September 4-6, 2013, Limerick, Ireland.* Abstracts. Ankara: Hacettepe University Department of Information Management..

Krovetz, R. (1993). Viewing morphology as an inference process. Robert Korfhage, Edie Rasmussen ve Peter Willett (Ed.), *SIGIR93 16th International ACM/SIGIR '93 Conference on Research and Development in Information Retrieval, June 27 - July 01, 1993, Pittsburgh, PA, USA içinde* (s. 191-202). Pittsburgh: ACM.

Kumaran, G. ve Allan, J. (2004). Text classification and named entities for new event detection. *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '04), July 25 - 29, 2004, Sheffield, United Kngdm* içinde (s. 297-304). New York, NY: ACM.

Kumaran, G. ve Allan, J. (2005). Using names and topics for new event detection. Raymond J. Mooney (Ed.), *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, October 6-8, 2005, Vancouver, British Columbia* içinde (s. 121-128). Morristown, N.J. : Association for Computing Linguistics.

Kurt, H. (2001). *On-line new event detection and tracking in a multi-resource environment*. Yayımlanmamış Yüksek Lisans Tezi. Ankara: Bilkent Üniversitesi Fen Bilimleri Enstitüsü. 10 Mayıs 2014 tarihinde http://www.thesis.bilkent.edu.tr/0001765.pdf adresinden erişildi.

Lakshmi, K. ve Mukherjee, S. (2007). Using cohesion-model for story link detection system. *IJCSNS International Journal of Computer Science and Network Security, 7*(3), 59-66.

Larose, D. T. (2005). *Discovering knowledge in data: an introduction to data mining*. USA:John Wiley & Sons.

Lavrenko, V. ve Croft, W. B. (2001). Relevance based language models. W. Bruce Croft (Ed.), *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, September 9-13, 2001, New Orleans, Louisiana, USA* içinde (s.120-127). New York, N.Y. : Association for Computing Machinery.

Lavrenko, V., Allan, J., DeGuzman, E., LaFlamme, D., Pollard, V. ve Thomas, S. (2002). Mitchell Marcus (Ed.), Relevance models for topic detection and tracking. *Proceedings of the second International Conference on Human Language Technology Research, March 24-27, 2002, San Diego, California* içinde (s. 115-121). San Francisco, Calif. : Morgan Kaufmann.

Lawrence, S. ve Giles, C.L. (1998). Searching the World Wide Web. *Science*, 280(3), 98–100.

Lee, J.H. (1995). Combining multiple evidence from different properties of weighting schemes. Edward A. Fox, Peter Ingwersen ve Raya Fidel (Ed.), *SIGIR '95: proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 9 - 13, 1995, Seattle, Washington* içinde (s.180-188). New York, NY: ACM.

Lee, S. ve Kim, H. J. (2008). News keyword extraction for topic tracking. *Networked Computing and Advanced Information Management.* 2, 554-559.

Leek, T., Schwartz, R. ve Sista, S. (2002). Probabilistic approaches to topic detection and tracking. James Allan (Ed.), *Topic Detection And Tracking* içinde (s. 67-83). Springer US.

Li, S., Lv, X., Li, Y. ve Shi, S. (2010a). Study on feature selection algorithm in topic tracking. Gang Kou (Ed.). *SEDM 2010: the 2nd International Conference on Software Engineering and Data Mining, 23-25 June 2010, Chengdu, China* içinde (s. 384-389). Piscataway, N.J. : IEEE.

Li, S., Lv, X., Zhou, Q. ve Shi, S. (2010b). Study on key technology of topic tracking based on VSM. *2010 IEEE International Conference on Information and Automation, June 20-23, 2010, Harbin, Heilongjiang, China* içinde (s. 2419-2423). Piscataway, N.J. : IEEE.

Liu, S. W. ve Chang, H. T. (2013). A topic detection and tracking system with TF-density. F. G. Gaol (Ed.). *Recent Progress in Data Engineering and Internet Technology* içinde (s. 115-120). Berlin, Heidelberg: Springer.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. L. M. L. Cam ve J. Neyman (Ed.), *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability* içinde (s. 281-297). Berkeley, Ca. ; Los Angeles, Ca. : University of California Press.

Makkonen, J., Ahonen, H. ve Salmenkivi, M. (2002). Applying semantic classes in event detection and tracking. Rajeev Sangal ve S.M. Bendre (Ed.), *Proceedings of International Conference on Natural Language Processing (ICON 2002), December 18-21, 2002, Mumbai, India* içinde (s. 175-183). New Delhi: Vikas Pub. House : Distributors, UBS Publishers' Distributors Ltd..

Makkonen, J., Ahonen, H. ve Salmenkivi, M. (2003). Topic detection and tracking with spatio-temporal evidence. Sebastiani Fabrizio (Ed.), *25th European Conference on IR Research, ECIR 2003, April 14–16, 2003, Pisa, Italy* içinde (s. 251-265). New York: Springer.

Maron, M. E. (1988). Probabilistic design principles for conventional and full-text retrieval systems. *Information Processing & Management,* 24(3), 249-255.

Maron, M.E. (1984). Probabilistic retrieval models. Brenda Dervin ve Melvin J. Voigt (Ed.), *Progress in Communication Sciences Vol. 5* içinde (s. 145-176). Norwood, NJ: Ablex.

Maron, M.E. ve Kuhns, J.L. (1960). On relevance, probabilistic indexing and information retrieval. *Journal of the Association for Computing Machinery, 7,* 216-244.

Martin, A. F., Doddington, G. R., Kamm, T., Ordowski, M. ve Przybocki, M. A. (1997). The DET curve in assessment of detection task performance. G. Kokkinakis, N. Fakotakis ve E. Dermatas (Ed.), *EUROSPEECH, ISCA, September, 1997, Rhodos* içinde (s. 1895-1898). Grenoble: ESCA.

Meadow, C. T. (1992). *Text Information Retrieval Systems*. San Diego: Academic Press.

McCallum, A., Nigam, K. ve Ungar, L. H. (2000). Efficient clustering of high-dimensional data sets with application to reference matching. Raghu Ramakrishnan, Sal Stolfo, Roberto Bayardo ve Ismail Parsa (Ed.), *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, Boston, MA, USA, August 20 - 23, 2000* içinde (s. 169-178). New York, NY, USA: Association for Computing Machinery.

Miller, D., Leek, T. ve Schwartz, R. (1999). A hidden markov model information retrieval system. *Proceedings on the 22nd Annual International ACM SIGIR Conference, 1999, Berkeley, California*, içinde (s. 214–221). New York, NY: ACM.

Mori, M., Miura, T. ve Shioya, I. (2006). Topic detection and tracking for news web pages. *WI '06 Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, December 8-22, 2006, Hong Kong, China* içinde (s. 338-342). Los Alamitos, Calif. : IEEE Computer Society.

Nomoto, T. (2010). Two-tier similarity model for story link detection. *CIKM '10 International Conference on Information and Knowledge Management, October 26 - 30, 2010, Toronto, ON, Canada* içinde (s. 789-798). New York, N.Y. : ACM Press.

Nowell, L. T., France, R. K., Hix, D., Heath, L. S. ve Fox, E. A. (1996). Visualizing search results: some alternatives to query-document similarity. Hans-Peter Frei, Donna Harman, Peter Schäuble ve Ross Wilkinson (Ed.), *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, August 18 - 22, 1996, Zurich, Switzerland* içinde (s. 67-75). New York: ACM.

Popovic, M. ve Willett P. (1992), The effectiveness of stemming for natural language access to Slovene textual data. *Journal of The American Society for Information Science*, *43*, 384-390.

Pena, J. M., Lozano, J. A. ve Larranaga, P. (1999). An empirical comparison of four initialization methods for the K-Means algorithm. *Pattern Recognition Letters*, *20*(10), 1027-1040.

Ponte, J. ve Croft, W. B. (1997). Text segmentation by topic. Carol Peters ve Costantino Thanos (Ed.), *Proceedings of the European Conference on Research and Advanced Technology for Digital Libraries (ECDL), September 1–3, 1997, Pisa, Italy* içinde (s.113-125). Berlin: Springer.

Ponte, J. ve Croft, W. B. (1998). A language modeling approach to information retrieval. W. Bruce Croft, Alistair Moffat, C.J. van Rijsbergen, Ross Wilkinson ve Justin Zobel (Ed.). *Proceedings on the 21st Annual International ACM SIGIR Conference, August 24 - 28, 1998, Melbourne, Australia* içinde (s. 275–281). New York, N.Y. : Association for Computing Machinery.

Qin, X. ve Zhang, Y. (2008). Improving the performance of topic tracking system by ensemble. *2008 International Conference* on *Computer Science and Software Engineering Volume 02, 12-14 December 2008, Wuhan, China* içinde (s. 316-320). IEEE Computer Society.

Qiu, J. ve Liao, L.J. (2008). Add temporal information to dependency structure language model for topic detection and tracking. *Machine Learning and Cybernetics, 12-15 July 2008, Kunming* içinde (s. 1575 – 1580). IEEE.

Qiu, J., Liao, L.J. ve Dong, X.J. (2008). Topic detection and tracking for Chinese news web pages. Cheolyoung Ock, JeongYong Byun, YuDe Bi ve Hongfei Lin (Ed.), *ALPIT 2008: Seventh International Conference on Advanced Language Processing and Web Information Technology: proceedings: 23-25 July 2008, Liaoning, China* içinde (s.114-120). Los Alamitos, CA: IEEE Computer Society.

Robertson, S.E. (1977). Theories and models in information retrieval. *Journal of Documentation, 33,* 126-148.

Robertson, S. E., Maron, M. E. ve Cooper, W. S. (1983). The unified probabilistic model for IR. Gerard Salton ve Hans-Jochen Schneider (Ed.). *Research and Development in Information Retrieval* içinde (s. 108-117). Berlin: Springer.

Salton, G. (1989). *Automatic Text Processing*. Reading, Mass: Addison-Wesley

Salton, G. ve Buckley, C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41, 288-97.

Salton, G. ve McGill, M.J. (1983). *Introduction to Modern Information Retrieval*. New York: McGraw Hill.

Salton, G., Wong, A. ve Yang, C. S. (1975), A vector space model for automatic indexing. *Communications of the ACM. 18*(11), 613–620.

Schultz, J. M. ve Liberman, M. (1999). Topic detection and tracking using idf-weighted cosine coefficient. *Proceedings of the DARPA broadcast news workshop, February 28-March 3, 1999, Herndon, Virginia* içinde (s. 189-192). San Francisco: Morgan Kaufmann.

Schultz, J. M. ve Liberman, M. Y. (2002). Towards a "Universal Dictionary" for multi-language information retrieval applications. James Allan (Ed.), *Topic Detection And Tracking* içinde (s. 225-241). US: Springer.

Sever, H. ve Bitirim Y. (2003). FindStem: Analysis and evaluation of a Turkish stemming algorithm. *Lecture Notes in Computer Science, 2857,* 238-251.

Shah, C. ve Eguchi, K. (2009). Improving document representation for story link detection by modeling term topicality. *Information and Media Technologies*, *4*(2), 433-441.

Shah, S. (2010). Caution: Reported trends in search query length may be misleading. *Search Engine Land*, 24 Nisan 2014 tarihinde http://searchengineland.com/caution-reported-trends-in-search-query-length-may-be-misleading-41641 adresinden erişildi.

Shah, C., Croft, W. B. ve Jensen, D. (2006). Representing documents with named entities for story link detection (SLD). Philip S. Yu, Vassilis J. Tsotras, Edward A. Fox ve Bing Liu (Ed.), *15th ACM Conference on Information and Knowledge Management (CIKM 2006), 6-11 November 2006, New York, USA* içinde (s. 868-869). New York, N.Y. : Association for Computing Machinery.

Sheridan, P. ve Ballerini, J.P., (1996). Experiments in multilingual information retrieval using the SPIDER System. Hans-Peter Frei, Donna Harman, Peter Schäuble ve Ross Wilkinson (Ed.), *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'96, August 18-22, 1996, Zurich, Switzerland (Special Issue of the SIGIR Forum)* içinde (s. 58–65). New York: ACM.

Solak, A. ve Can, F. (1994). Effects of stemming on Turkish text retrieval. *Proceedings of the Ninth Int. Symp. on Computer and Information Sciences (ISCIS '94), November 1994, Antalya, Turkey* içinde (s. 49-56). Antalya:IEEE.

Song, F. ve Croft, W. B. (1999). A general language model for information retrieval. *SIGIR '99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in InformationRetrieval, August 15-19, 1999, Berkeley, CA, USA* içinde (s. 279–280). New York, NY: ACM.

Song, M., Song, I. Y., Hu, X. ve Allen, R. B. (2007). Integration of association rules and ontologies for semantic query expansion. *Data & Knowledge Engineering*, *63*(1), 63-75.

Soydal, İ. ve Al, U. (2014). *Türkçe Haber Benzerliklerinin Belirlenmesinde Varlık İsimlerinin Hikâye Bağlantı Algılama Görevinin Başarımına Etkisi*. TÜBİTAK Sosyal Bilimler Araştırma Grubu - Proje No: SOBAG 111K030. Hacettepe Üniversitesi Bilgi ve Belge Yönetimi Bölümü. Ankara, 2014. (iv, 47 s.).

Sparck Jones, K., Walker, S. ve Robertson, S.E. (2000). A probabilistic model of information retrieval: development and comparative experiments. *Information Processing & Management*, *36*(2), 809-840.

Spink, A., Wolfram, D., Jansen, M. B. ve Saracevic, T. (2001). Searching the web: The public and their queries. *Journal of the American Society for Information Science and Technology*, *52*(3), 226-234.

Steinbach, M., Karypis, G. ve Kumar, V. (2000). A comparison of document clustering techniques. Marko Grobelnik, Dunja Mladenic ve Natasa Milic-Frayling (Ed.). *Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 20-23, 2000, Boston, MA, USA* içinde (s. 525-526).

TDT. (2002). *The 2002 topic detection and tracking (TDT2002) task definition and evaluation plan. Technical Report Version 1.1*, National Institute of Standards and Technology, 2002.

Thompson, K.C. ve Callan, J. (2005). Query expansion using random walk models. Otthein Herzog, Hans-Jörg Schek, Norbert Fuhr, Abdur Chowdhury ve Wilfried Teiken (Ed.), *Proceedings of the 2005 ACM CIKM International Conference on Information and Knowledge Management, October 31 - November 5, 2005, Bremen, Germany* içinde (s. 704-711). ACM.

Tonta, Y. (1995). Bilgi Erişim Sistemleri. *Türk Kütüphaneciliği, 9*(3), 302-314.

Tonta, Y., Bitirim. Y. ve Sever. H. (2002). *Türkçe Arama Motorlarında Performans Degerlendirme,* Ankara: Total Bilişim Ltd. Şti..

Torunoglu, D., Cakirman, E., Ganiz, M. C., Akyokus, S. ve Gurbuz, M. Z. (2011). Analysis of preprocessing methods on classification of Turkish texts. *Innovations in Intelligent Systems and Applications (INISTA), 2011 International Symposium, 5-18 June 2011, Istanbul, Turkey* içinde (s. 112-117). Piscataway, NJ: IEEE.

Tunali, V. ve Bilgin, T. (2012). Examining the impact of stemming on clustering Turkish texts. *Innovations in Intelligent Systems and Applications (INISTA), 2012 International Symposium, 2-4 July 2012, Trabzon, Turkey* içinde (s. 1-4). Piscataway, NJ: IEEE.

Viermetz, M., Skubacz, M., Ziegler, C. N. ve Seipel, D. (2008). Tracking topic evolution in news environments. *10th IEEE International Conference on E-Commerce Technology (CEC 2008) / 5th IEEE International Conference on Enterprise Computing, E-Commerce and E-Services (EEE 2008), July 21-14, 2008, Washington, DC, USA* içinde (s. 215-220). Los Alamitos, Calif. : IEEE Computer Society.

Voorbij, H.J. (1999). Searching scientific information on the Internet: A Dutch academic user survey. *Journal of the American Society for Information Science, 50*(7), 598–615.

Vural, A. (2002). On-line new event detection and clustering using the concepts of the cover coefficient-based clustering methodology. Yayımlanmamış Yüksek Lisans Tezi. Bilkent Üniversitesi Fen Bilimleri Enstitüsü. Ankara. 10 Mayıs 2014 tarihinde http://www.cs.bilkent.edu.tr/tech-reports/2002/BU-CE-0218.pdf adresinden erişildi.

Walls, F., Schwartz, R., Jin, H. ve Sista, S. (1999). Probabilistic models for topic detection and tracking. *Acoustics, Speech, and Signal Processing, 1999. Proceedings. Vol 1, 1999 IEEE International Conference, 15-19 Mar 1999, Phoenix, AZ* içinde (s. 521-524). New York, N.Y. : IEEE.

Wang, C., Zhang, M., Ru, L. ve Ma, S. (2008). An automatic online news topic keyphrase extraction system. Lakhmi Jain ve diğerleri (Ed.), *2008 IEEE / WIC / ACM International Conference on Web Intelligence, WI 2008, 9-12 December 2008, Sydney, NSW, Australia, Main Conference Proceedings* içinde (s. 214-219). Los Alamitos, Calif. : IEEE Computer Society.

Wolfram, D. (1999). Term co-occurrence in Internet search engine queries: An analysis of the Excite data set. *Canadian Journal of Information and Library Science*, 24(2/3): 12-33.

Xiaowei, W., Longbin, J. ve Jialin, M. (2008). Use of NER information for improved topic tracking. Jeng-Shyang Pan, Ajith Abraham ve Chin-Chen Chang (Ed.), *Eighth International Conference on Intelligent Systems Design and Applications, ISDA 2008, 26-28 November 2008, Kaohsiung, Taiwan* içinde (s. 165-170). Los Alamitos, Calif. : IEEE Computer Society.

Xu, J. ve Croft, W. B. (2000). Improving the effectiveness of information retrieval with local context analysis. *ACM Transactions on Information Systems (TOIS), 18*(1), 79-112.

Xu, J. ve Croft, W. B. (1996). Query expansion using local and global document analysis. Hans-Peter Frei, Donna Harman, Peter Schäuble ve Ross Wilkinson (Ed.), *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'96, August 18-22, 1996, Zurich, Switzerland (Special Issue of the SIGIR Forum)* içinde (s. 4-11). New York: ACM.

Xu, R. ve Wunsch, D. (2005). Survey of clustering algorithms. *Neural Networks*, IEEE Transactions on, *16*(3), 645-678.

Yamron, J. P., Carp, I., Gillick, L., Lowe, S. ve Van Mulbregt, P. (1997). Event tracking and text segmentation via hidden markov models. Sadaoki Furui, B.H. Juan ve Wu Chou (Ed.), *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop, 14-17 Dec 1997, Santa Barbara, CA* içinde (s. 519-526). Piscataway, NJ: IEEE.

Yamron, J. P., Carp, I., Gillick, L., Lowe, S. ve Van Mulbregt, P. (1998). A hidden Markov model approach to text segmentation and event tracking. *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech, and Signal Processing, May 12-15, 1998, Seattle, Washington (USA)* içinde (s. 333-336). Piscataway, NJ: IEEE Service Center.

Yamron, J., Knecht, S. ve Mulbregt, P. (2000). Dragon's tracking and detection systems for the TDT2000 evaluation, *Proceedings of Topic Detection and Tracking Workshop, 8-11 February 1998, Lansdowne, Virginia* içinde (s. 75-80).

Yang, Y., Ault, T., Pierce, T. ve Lattimer, C. W. (2000). Improving text categorization methods for event tracking. Nicholas J. Belkin, Peter Ingwersen ve Mun-Kew Leong (Ed.), *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM SIGIR, July 24-28, 2000, Athens, Greece* içinde (s. 65-72). New York, N.Y. : Association for Computing Machinery.

Yang, Y., Carbonell, J., Brown, R., Lafferty, J., Pierce, T. ve Ault, T. (2002). Multi-strategy learning for topic detection and tracking. James Allan (Ed.), *Topic Detection and Tracking* içinde (s. 85-114). USA:Springer.

Zhang, X., Guo, Z. ve Li, B. (2009, May). An effective algorithm of news topic tracking. *Proceedings of the 2009 WRI Global Congress on Intelligent Systems (Volume:3), 19-21 May 2009, Xiamen, China* içinde (s. 510-513). Los Alamitos, Calif. : IEEE Computer Society.

ÖZGEÇMİŞ

Kişisel Bilgiler

Adı Soyadı : Güven KÖSE
Doğum Yeri ve Tarihi : Ardanuç – 23.04.1975

Eğitim Durumu

Lisans Öğrenimi : Gazi Üniversitesi, Teknik Eğitim Fakültesi, Bilgisayar Sistemleri Öğretmenliği
Yüksek Lisans Öğrenimi : Başkent Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği
Bildiği Yabancı Diller : İngilizce
Bilimsel Faaliyetleri : Sever, H., Akal, F. ve Köse, G. (2007). Kavram Tabanlı Bilgi Geri Getirim Yaklaşımı. Bilgi Dünyası. 8(1): 49-75.

Sever, H. ve Köse, G. (2006). Üst Arama Motorları. Türkiye Bilişim Ansiklopedisi, T.Ören, T. Üney ve R. Çölkesen / İstanbul: Papatya Yayıncılık, Ankara.

Köse, G. and Ahmadlouei, H. (2013). Supervised News Classification Based on Large-scale News Corpus, 4th International Symposium on Information Management in a Changing World, Semtember 4-6, 2013, Limerick Institute of Technology, Springer sponsors.

Köse, G., Tonta, Y., Polatkan, A.C. and Ahmadlouei, H. (2013). Story Link Detection in Turkish Corpus, The 2013 IEEE/WIC/ACM International Conference on Web Intelligence, Nov. 17-20, 2013 Atlanta GA USA.

İş Deneyimi

Stajlar : Gazi Üniversitesi, Teknik Eğitim Fakültesi
Projeler : Evliya Çelebi Coğrafi Bilgi Sistemi Çekirdeği (Araştırmacı, TUBITAK)
DrCAD Tıbbi Karar Destek Sistemi (Danışman, TUBITAK)
MANTAM Türkçe Arama Motoru (Danışman, TUBITAK)
Türkçe Haber Benzerliklerinin Belirlenmesinde Varlık İsimlerinin Hikâye Bağlantı Algılama Görevinin Başarımına Etkisi (Doktora Öğrencisi, TUBITAK - 111K030)
Çalıştığı Kurumlar : Başkent Üniversitesi (1999-2009)
Hacettepe Üniversitesi (2009-2012)
Mantis Yaz. Dan. Ltd. Şti. (2012-Devam Ediyor)

İletişim

E-Posta Adresi : guvenkose@gmail.com

Tarih : 05.06.2014